AI-Empowered Person Re-Identification Based on Skeleton Data

Haocong Rao

College of Computing and Data Science

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2025

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

21/01/2025

Date

Haocong Rao

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

21/01/2025

.....

Date

.....

Prof. Chunyan Miao

Authorship Attribution Statement

This thesis contains material from two paper accepted at conferences and two paper published in the following peer-reviewed journals in which I am listed as an author.

Chapter 4 is published as Haocong Rao and Chunyan Miao. "SimMC: Simple Masked Contrastive Learning of Skeleton Representations for Unsupervised Person Re-Identification," In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 2022, pp. 1290-1297. DOI: 10.24963/ijcai.2022/180.

The contributions of the co-authors are as follows:

- Prof. Miao pointed out the overall research direction and provided valuable constructive suggestions on the work.
- I designed the model, developed the source codes, conducted the experiments/analyses, and drafted the manuscript at the LILY Research Center and College of Computing and Data Science.
- Prof. Miao provided helpful research materials and discussions, as well as helped improve the idea and revise the manuscript.

Chapter 5 is published as Haocong Rao, Cyril Leung, and Chunyan Miao, "Hierarchical Skeleton Meta-Prototype Contrastive Learning with Hard Skeleton Mining for Unsupervised Person Re-Identification," *International Journal of Computer Vision (IJCV)*, vol. 132, pp. 1–23, 2023. DOI: 10.1007/s11263-023-01864-0.

The contributions of the co-authors are as follows:

- Prof. Miao, Prof. Leung, and I discussed the initial research direction.
- I designed the model, developed the source codes, conducted the experiments/analyses, and drafted the manuscript at the LILY Research Center and College of Computing and Data Science.
- Prof. Miao and Prof. Leung provided useful comments and discussions on the problem, techniques, experiments, mathematical modeling, and revised/edited the manuscript.

Chapter 6 is published as Haocong Rao and Chunyan Miao. "TranSG: Transformer-Based Skeleton Graph Prototype Contrastive Learning with Structure-Trajectory Prompted Reconstruction for Person Re-Identification," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22118-22128. DOI: 10.1109/CVPR52729.2023.02118.

The contributions of the co-authors are as follows:

- Prof. Miao and I discussed the initial research direction and provided valuable comments on the idea.
- I designed the model, developed the source codes, conducted the experiments/analyses, and drafted the manuscript at the LILY Research Center and College of Computing and Data Science.
- Prof. Miao provided helpful research materials and discussions, as well as helped improve the idea and revise the manuscript.

Chapter 7 is published as Haocong Rao, Yuan Li, and Chunyan Miao, "Revisiting k-Reciprocal Distance Re-ranking for Skeleton-Based Person Re-Identification," *IEEE Signal Processing Letters (SPL)*, vol. 29, pp. 2103–2107, 2022. DOI: 10.1109/LSP.2022.3212634.

The contributions of the co-authors are as follows:

- Prof. Miao, Prof. Li, and I discussed the initial research direction.
- I designed the model, developed the source codes, conducted the experiments/analyses, and drafted the manuscript at the LILY Research Center and College of Computing and Data Science.
- Prof. Miao and Prof. Li offered useful comments on the problem, techniques and revised/edited the manuscript.
- Prof. Miao provided helpful reading materials.

21/01/2025

Date

Haocong Rao

Acknowledgements

I wish to express my greatest gratitude to my main supervisor Prof. Chunyan Miao and co-supervisor Prof. Cyril Leung for their patient guidance, profound encouragement, and kind help during my PhD study. Not only do they provide many invaluable suggestions that significantly inspire my research work, but they also generously share precious insights that illuminate the paths of my career and life. Furthermore, I would like to extend my heartfelt gratitude to my Thesis Advisory Committee (TAC) members (Prof. Guosheng Lin and Prof. Huaxiong Wang) and Qualifying Examination (QE) committee members (Prof. Chen Change Loy, Prof. Ziwei Liu, and Prof. Huaxiong Wang), for their engaging academic discussions and professional guidance. It is also my immense honor to meet and learn from many outstanding professors and teachers, and I hope to wholeheartedly thank them as well, including: My course teachers (Prof. Tat Jen Cham, Dr. SoH Leong Ping Alvin, Prof. Jianming Zheng, Prof. Xiaosheng Qin, Prof. Feng Zhu, Prof. Newton Fernando, Prof. Ying He, Prof. Boyang Li, Dr. Sujata Surinder Kathpalia), who taught me a broad range of knowledge and enriched my professional skills; My previous academic mentors (Prof. Xiping Hu, Prof. Victor C. M. Leung, Prof. Siqi Wang, Prof. Mingkui Tan, Prof. Bin Hu, Prof. Jun Cheng), who constantly supported and helped me to improve the academic ability and build a solid foundation for my PhD study.

I would like to sincerely thank my colleagues and collaborators at the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), including Bo Huang, Chang Liu, Di Wang, Elsie Sim Lee Boon, Fanglin Zhu, Hana Ahmad, Hangwei Qian, Hao Wang, Haochen Li, He Zhao, Hongyu Qiu, Hongyu Zhou, Huiguo Zhang, Jessica Hon-Chan, Jiaxi Gao, Jiazheng Jing, Jun Lin, Leung Jonathan Cyril, Liang Zhang, Lim Hiong Loo, Lim Su Fang, Ming Chen, Minlin Zeng, Pengcheng Wu, Ping Chen, Robin Chan Chung Leung, Shengfei Lyu, Shibo Feng, Shuaicheng Niu, Sixuan Du, Siyuan Liu, Tan Toh Hsiang Benny, Tian Tian, Tiantong Wang, Tianyi Wang, Tong Zhang, Wanjin Feng, Wei Wang, Xiao Yang,

Xiaoqi Qiu, Xiaoxiong Zhang, Xinjia Yu, Xu Guo, Xuejiao Zhao, Yaming Zhang, Yanci Zhang, Yang Qiu, Yidan Hu, Yige Xu, Yinan Zhang, Yixin Zhang, Yong Liu, Yonghui Xu, Yongjie Wang, Yu Han, Yuanyuan Chen, Yue Yu, Zhengxiang Pan, Zhiwei Zeng, Zhixiang Su, Zijie Wang, Zuhan Meng and many others, for their help and suggestions during the times when I encounter difficulties. I really enjoy the time communicating, studying, and working with them, who made my stay in Singapore feel just like being at home.

I also hope to express my tremendous thanks to all my friends, including but not limited to Bowang Li, Bo Zhang, Changmeng Zheng, Chunyun Chen, Dongyuan Shi, Fengming Liu, Guofu Zhu, Guoji Fu, Haifeng Lu, Haozhe Ma, Hengze You, Jiaze Li, Jiahui Zhang, Jiayun Luo, Jiaze Li, Jiyuan Shen, Junqiao Fan, Junqin Lin, Kaijun Shen, Linan He, Luo Zhang, Luoying Hao, Luqman Alka, Muming Lian, Nanhong Chen, Peijun Bao, Qian Dong, Ruihang Wang, Shakya Manoj, Shihao Xu, Shuai Wang, Siyu Zhu, Tianshu Gao, Wei Gao, Wei Nong, Xiaocong Luo, Xiaohao Lin, Xiaoli Tang, Xingxuan Li, Xuhui Zhou, Youming Fan, Yuanjing Chen, Yuehua Li, Zejin Chen, Zhengding Luo, Zhijuan Shen, Zhiwei Cao, and Zongkai Li, for the memorable moments in my life as well as their enduring care, encouragement, and assistance.

Last but not least, I would like to thank my family for their unconditional love and unwavering support, always encouraging me to pursue my life goals. To my first love, my soulmate and my girlfriend, Ms. Yunxi Zhuang, I hope to express my deepest and sincerest gratitude, for her unparalleled love and patience, for believing in me more than I do, for always together to fight through hardships. It was my fortune for meeting her, and it was also my fortune to be with her at the hardest time of our lives. She had the kindest soul and purest heart with the strongest will that I have ever known in this world, and I believe she now lives the happiest and freest life beyond this world. The courage she gave me will continuously lights my path—step by step, I'll honor the promise we made.

"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity."

—Charles Dickens

"博学之，审问之，慎思之，明辨之，笃行之。"

—《礼记 · 中庸》

To my dear family

Abstract

Person re-identification (re-ID) plays a pivotal role in many safety-critical and daily applications such as intelligent video surveillance and human tracking. Despite great effort, conventional appearance-based methods using images or videos are usually computationally expensive and vulnerable to appearance and environmental variations. With the popularization of low-cost and accurate skeleton-tracking sensors, person re-ID via *skeleton data* is becoming an important emerging topic with prominent advantages. It focuses on modeling of unique body structure and motion patterns based on key body-joint positions, which typically enjoys smaller data inputs and model sizes, more robust performance under shape and view variations, as well as better privacy protection without using any visual appearances.

This thesis focuses on effectively modeling and learning discriminative body structure features and motion patterns from *skeleton data* for person re-ID. We design deep neural network based AI models and frameworks to perform *automatic* skeleton representation learning for person re-ID, so as to address the primary limitation of most previous methods that rely on hand-crafted skeleton descriptors and often fail to capture latent features beyond domain knowledge. In particular, we concentrate on and address several key challenges in this area. Firstly, most existing methods require massive skeleton labels for model training, while the collection of skeletal annotations is usually labor-expensive and unavailable in practice. Such label dependency decreases their applicability and generality in real-world open scenarios. Secondly, existing endeavors typically learn skeleton features from a single level (*e.g.*, body joint level), while this intrinsically limits their ability to exploit valuable *hierarchical* skeleton features and higher level patterns. It is still an open challenge to model discriminative body structure and motion features at *different levels* for person re-ID. The third challenge is exploiting latent *relations* between different body joints or components to capture more recognizable patterns. The way to model valuable body structural and actional relations from skeleton graph representations remains to be an unresolved issue. Another crucial challenge is

devising *generic auxiliary* tasks or techniques (*e.g.*, feature re-ranking) that can be flexibly applied to different models to enhance their person re-ID performance.

To this end, this thesis explores AI-empowered solutions to specifically address the aforementioned challenges. Firstly, to learn effective skeleton representations *without* labels, we propose a generic unsupervised Simple Masked Contrastive learning (SimMC) framework for person re-ID. SimMC leverages randomly-masked skeleton subsequences to mine the most representative skeleton features (termed *prototypes*), and capture valuable intra-sequence skeleton semantics via contrasting motion continuity between subsequences. Empirical evaluations demonstrate effectiveness, efficiency, and generality of SimMC on multiple benchmark datasets.

To solve the second challenge of multi-level skeleton modeling, we propose a Hierarchical skeleton Meta-Prototype Contrastive learning (Hi-MPC) approach. Hi-MPC constructs *coarse-to-fine* body representations at different levels, and exploits them to learn most typical skeleton features (termed meta-prototypes) from *multiple* contrastive representation spaces, which improves the robustness and consistency of previous skeleton prototype learning. A hard skeleton mining mechanism is devised to learn skeleton importance and more valuable patterns. Extensive experiments on five datasets demonstrate its state-of-the-art performance and scalability.

To address the third problem to fully capture latent body relations, we propose a Transformer-based Skeleton Graph prototype contrastive learning (TranSG) paradigm. We devise the skeleton graph transformer to *simultaneously* capture structural and actional body relations, and propose the graph prototype contrastive learning to capture more reliable prototypes and class-related semantics. The graph structure-trajectory prompted reconstruction is proposed to learn general high-level skeleton semantics. Extensive experiments on five benchmark datasets demonstrate the effectiveness of TranSG and its high generality under different scenarios.

Lastly, to achieve general model enhancement, we propose a skeleton feature re-ranking technique to improve different models. We devise the skeleton sequence pooling to aggregate the most salient skeleton features. Then, both k-reciprocal and Euclidean distance are fused to integrate neighbors' context for more reliable feature re-ranking. A context-based voting scheme is further proposed to better select Rank-1 candidate for person re-ID. Empirical and qualitative evaluations on multiple datasets show the efficacy and generality of the proposed technique.

Contents

List of Figures

List of Tables

Acronyms

Person re-ID	Person re-IDentification
Hi-MPC	Hierarchical skeleton Meta-Prototype Contrastive learning
HSM	Hard Skeleton Mining
DNN	Deep Neural Networks
MSMR	Multi-level Skeleton Meta-Representation
D_{13}	13 skeleton Descriptors in [1]
D_{16}	16 skeleton Descriptors in [2]
CNN	Convolutional Neural Networks
LSTM	Long Short-Term Memory
AGE	Attention-based Gait Encoding [3]
SGELA	Self-supervised Gait Encoding with Locality-Awareness [4]
SimMC	Simple Masked Contrastive learning framework [5]
NCE	Noise-Contrastive Estimation
CPC	Contrastive Predictive Coding
TAP	Temporal Average Pooling
DBSCAN	Density-Based Spatial Clustering of Applications with Noise [6]
MLP	Multi-Layer Perceptron
KGBD	Kinect Gait Biometry Dataset
BIWI	BIWI RGBD-ID Dataset
KS20	KS20 VisLab Multi-View Kinect Skeleton Dataset
IAS	IAS-Lab RGBD-ID Dataset
CASIA-B	Chinese Academy of Sciences, the Institute of Automation (CASIA) Gait Database B
IDs	IDentities
RVE	Random View Evaluation
CVE	Cross-View Evaluation

CMC	Cumulative Matching Characteristic
R_1	Rank-1 accuracy
R_5	Rank-5 accuracy
R_{10}	Rank-10 accuracy
mAP	mean Average Precision
MG-SCR	Multi-level Graph encoding with Structural-Collaborative Relation learning [7]
SM-SGE	Self-supervised Multi-scale Skeleton Graph Encoding [8]
BIWI-S	BIWI Still testing (probe) set
BIWI-W	BIWI Walking testing (probe) set
IAS-A	IAS-Lab testing (probe) set A
IAS-B	IAS-Lab testing (probe) set B
LMNN	Large Margin Nearest Neighbor (classification) [9]
ITML	Information-Theoretic Metric Learning [10]
ELF	Ensemble of Localized Features [11]
SDALF	Symmetry-Driven Accumulation of Local Features [12]
MLR	Metric Learning to Rank [13]
DPC	Direct Prototype Contrastive learning
MPC	Meta-Prototype Contrastive learning
Hi	Hierarchical skeleton representations
t-SNE	t-distributed Stochastic Neighbor Embedding
MSMT17	Multi-Scene Multi-Time person re-identification dataset
MPC	Masked Prototype Contrastive learning
MIC	Masked Intra-sequence Contrastive learning
DF	Direct supervised Fine-tuning
NPC	Naïve Prototype Contrastive learning
SSP	Skeleton Sequence Pooling
AP	Average Pooling
MP	Max Pooling
k-NN	k-Nearest Neighbors
k-RR	k-Reciprocal distance Re-Ranking
FDR	Fused Distance based Re-ranking
RV	context-based Rank-1 Voting
Add.	Addition
Concat.	Concatenation

TranSG	Transformer-based Skeleton Graph prototype contrastive learning [14]
SGT	Skeleton Graph Transformer
GPC	Graph Prototype Contrastive learning
STPR	Structure-Trajectory Prompted Reconstruction
FR	Full-Relation
FFN	Feed Forward Network
EM	Expectation-Maximization
DS	Direct Supervised learning
GEI	Gait Energy Image
GEV	Gait Energy Volume
PCM	Point Cloud Matching
ICL	Instance-wise Contrastive Learning
LLMs	Large Language Models

Chapter 1

Introduction

1.1 Background and Motivation

Person re-identification (re-ID) is a pattern recognition task of retrieving and matching a certain pedestrian across different views, occasions, and scenarios. When given a walking sequence such as a surveillance video containing the person-of-interest, the target of person re-ID is to query (*e.g.*, match and predict) the correct identity of this person based on the existing records in the database [15]. Due to the increasing demand of public safety and emerging large-scale camera networks, person re-ID has played a pivotal role in many daily and safe-critical applications such as intelligent video surveillance, security authentication, human tracking, healthcare monitoring, and robotics [16–20]. The estimated global market size of this technology and its related domains (*e.g.*, digital identity authentication) was US$27.9 billion in 2022 and is expected to reach US$70.7 billion by 2027 [21], highlighting its vast application prospects.

In the past few years, with the surging popularity of low-cost, non-intrusive, and accurate skeleton-tracking sensors (*e.g.*, Kinect [22]), person re-ID based on skeleton data has attracted increasing attention in both academia and industry [23–26]. In the task of skeleton-based person re-ID, when given a skeleton sequence (which is captured from sensors or estimated by models) containing the person-of-the-interest, the target of model is to query the identity of this person from the database (illustrated in Fig. 1.1). Compared with conventional image or video based methods that typically rely on visual features such as human silhouettes and

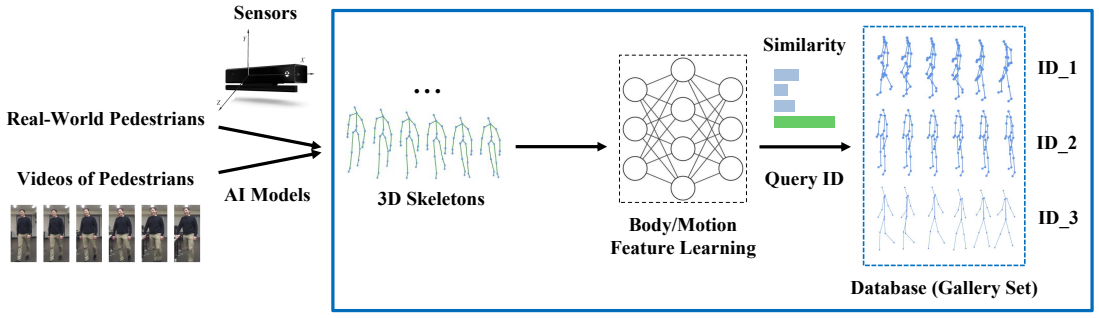


FIGURE 1.1: General process of skeleton-based person re-identification (re-ID). The input skeleton data of pedestrians are captured from depth sensors or estimated by pose estimation AI models, while the main task (shown in the blue box) is to learn effective representation of a skeleton sequence to query its identity (ID) via matching with existing sequences from the gallery set.

appearances (*e.g.*, faces) for recognition [27–29], skeleton-based methods leverage 3D positions of key body joints to characterize discriminative structural and motion features of human body, which could enjoy many advantages: (1) *Smaller data inputs* compared with image/video data; (2) *Lighter models* to process and learn skeletons than image-based models; (3) *Better privacy protection* without using any visual appearances; (4) *Higher robustness* under variations of scales, views or other external factors [30]. For example, many appearance-based methods and their texture-based body representations are sensitive to environmental factors such as background clutters, occlusion, and illumination [31], and usually require high-quality (*e.g.*, high-resolution) images for model leaning. Such image data often contain large-scale pixels that are computationally expensive in practice. By contrast, based on highly concise skeletal body representations, the design of artificial intelligence (AI) models (*e.g.*, deep learning based architectures) for skeleton-based person re-ID can be more efficient with significantly less parameters and computational complexity [5]. These advantages enable them to be potentially applied to more real-world scenarios with lower deployment cost. For instance, the small input data and low resource requirement allow skeleton-based person re-ID models to be potentially integrated into different portable RGB-D devices including Intel Realsense [32] and Apple Vision Pro to perform diverse identity-related pattern recognition tasks. In healthcare areas, the 3D skeleton data can be potentially applied to construct *efficient* human body representations to learn light-weight AI models for different tasks such as identity-aware health monitoring, gait analysis, disease diagnosis, etc [33].

To perform person re-ID via skeleton data, existing endeavors typically model

skeleton features by two groups of methods: (1) *Skeleton descriptor based methods* [2, 23, 24, 34], which manually extract certain anthropometric, geometric, and gait attributes of human body from skeleton data. However, these handcrafted methods usually require domain knowledge such as human anatomy [35], and cannot fully mine underlying features beyond human cognition; (2) *Deep neural network based methods* [3, 4, 25], which usually leverage convolutional neural networks (CNN) or long short-term memory (LSTM) [36] to learn skeleton representations with sequences of raw body-joint positions or pose descriptors (*e.g.*, motion angles). In practical terms, these methods usually require massive labeled skeleton data of pre-defined classes to either train the model from scratch (*e.g.*, supervised models [7, 25]) or fine-tune the pre-trained skeleton representations (*e.g.*, self-supervised models [3, 4, 8]) to classify the known identities. However, for existing supervised methods, they lack the flexibility to learn general and representative skeleton features that can re-identify different pedestrians under the unavailability of labels, which limits its application in many real-world scenarios. For existing self-supervised models, their performance is often unsatisfactory, and they typically necessitate more intricate design in terms of architectures and training process (*e.g.*, two-stage training combined with label-based fine-tuning).

1.2 Research Problem and Challenges

The **core research problem** of this thesis can be summarized as *"How to effectively model and learn skeleton data to perform person re-ID?"*, more specifically and essentially, *"How to effectively model and learn discriminative body structure features and motion patterns based on 3D skeleton data to identify different persons?"*. To this end, the focused **overall goal** is to devise deep neural network based AI models and frameworks to perform *automatic* effective skeleton representation learning and progressively improve their performance. Conventional skeleton-based methods [2, 23, 24, 34] manually design skeleton descriptors or pose features for human recognition, while their performance are limited by existing domain knowledge (*e.g.*, human anatomy [35]) and may ignore some latent representative features beyond human cognition. For example, there usually exist *implicit* correlations between pose features (*e.g.*, joint angles) of different skeletons

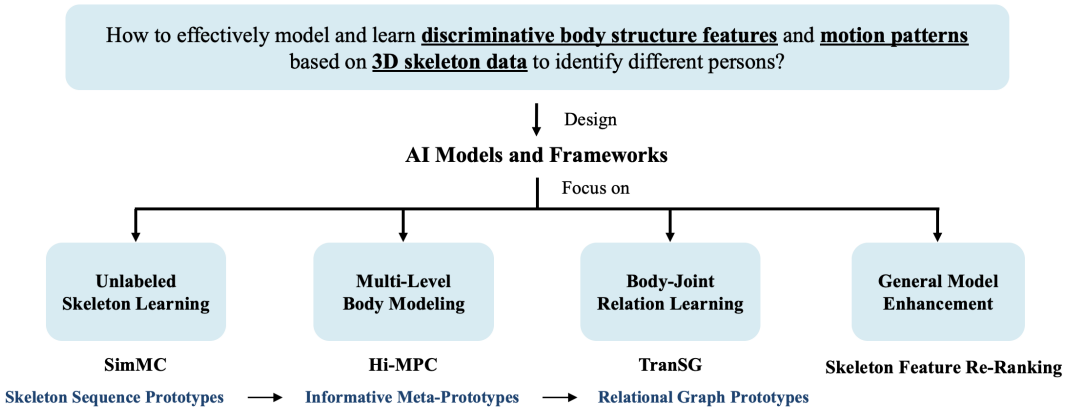


FIGURE 1.2: Overview of overall target and four key challenges in skeleton-based person re-ID. This thesis proposes AI-empowered solutions (SimMC, Hi-MPC, TranSG, Skeleton Feature Re-Ranking) to respectively address these challenges, and progressively improve the skeleton prototype learning method to achieve more effective skeleton-based person re-ID.

or continuous gait cycles, which can be encoded by sequence-learning AI models such as LSTM to capture unique and recognizable walking patterns (*i.e.*, gait [37]). Thus, how to devise effective AI models, including neural network architectures and objective functions, to automatically mine these valuable features is the critical target of this area. On top of this goal, we further concentrate on **four key challenges** to achieve more effective and efficient skeleton-based person re-ID as outlined below (illustrated in Fig. 1.2):

The first challenge is learning effective skeleton representations in an *unsupervised* manner (defined as "*unlabeled skeleton learning*"). Previous skeleton learning paradigms often require manually-annotated skeleton data to train or fine-tune the model [3, 7, 25], while this could reduce their generality for real-world scenarios, as the collection of massive labels is usually labor-expensive and unavailable in practice. In addition, using the pre-defined close-set classes weakens the flexibility of the model to be applied to more open scenarios that contain unknown identities. In this context, how to fully exploit *unlabeled* skeleton data to learn representative skeleton features and general skeleton semantics (*e.g.*, motion patterns and continuity) for person re-ID is a key research aspect of the thesis.

The second challenge is modeling body structure and motion features from *different levels* (defined as "*multi-level body modeling*"), as a comprehensive and general pre-modeling of skeleton data is often crucial for feature representation learning and can facilitate model training [8, 38]. Existing methods typically learn skeleton

features from a single level (*e.g.*, body joint level [3, 4]). However, this intrinsically limits their ability to exploit more valuable hierarchical features and high-level patterns from skeleton data. Besides, since different-level skeleton representations usually possess different contributions to the model performance, it is also crucial to integrate their importance into skeleton modeling and model training. For instance, some skeleton representations are more difficult to be classified to the correct identity but can provide more informative clues for discriminating different patterns [39], thus deserving more attention in skeleton learning.

The third challenge is capturing latent *relations* among different body joints or components (defined as "*body-joint relation learning*"). Existing methods typically encode pairwise joint distances (*e.g.*, limb lengths) or the trajectory of body-joint positions into a feature vector for modeling skeleton dynamics. However, they rarely explore latent relations between different body joints or components, thus ignoring valuable structural and actional information of human body. Take people's walking for example, adjacent body joints such as "knee", "foot", and collaborative limbs like "arm", "leg", usually exhibit different internal relational patterns during movement, which could carry unique relational information and recognizable walking patterns for person re-ID [37].

The fourth challenge is devising and combining *generic auxiliary* techniques or tasks, such as features post-processing techniques (*e.g.*, feature re-ranking) to improve the person re-ID performance of different models (defined as "*general model enhancement*"). As the skeleton-based person re-ID can be viewed as a retrieval problem, and its performance could be influenced by different distance metrics and the ranking results of feature representations [5], it is beneficial to devise feature re-ranking techniques to synergize different metrics and contexts to improve the ranking quality and model performance. On the other hand, since capturing general effective and discriminative skeleton semantics is a pivotal part of skeleton-based pattern recognition models [4, 40], how to effectively exploit key properties of skeleton sequences, such as motion consistency and pattern invariance [4, 5], to learn general and identity-specific semantics is still an open challenge.

1.3 Contributions

This section outlines the main contributions of the thesis, while the elaborated contributions of each proposed skeleton-based person re-ID model or framework will be presented in the respective chapters. Focusing on the four challenges (identified in Sec. 1.2) confronted by existing skeleton-based person re-ID methods, our contributions can be summarized as follows:

- To address the first challenge of unlabeled skeleton learning, we propose a generic **unsupervised** skeleton feature learning framework, named *Simple Masked Contrastive learning (SimMC)* framework (see Chapter 4), which contrasts the typical features and inherent relationships of *masked* skeleton sequences to learn effective skeleton representations **without using any label** for person re-ID. SimMC consists of two key components, *masked prototype contrastive learning (MPC)* and *masked intra-sequence contrastive learning (MIC)*. MPC exploits the most representative features (termed *prototypes*) of *masked* skeleton sequences to learn discriminative representations of unlabeled skeleton data, and MIC performs intra-sequence feature learning by contrasting different subsequences and their motion continuity to encourage more effective skeleton semantics learning for person re-ID. Empirical evaluations demonstrate both effectiveness and efficiency of SimMC on multiple benchmark datasets under different scenarios. SimMC achieves highly competitive performance with a lightweight Siamese architecture and also possesses good theoretical interpretability. It can also serve as a generic contrastive paradigm to fine-tune and boost existing skeleton representations.

- To solve the second challenge of multi-level body modeling, we devise an unsupervised *Hierarchical skeleton Meta-Prototype Contrastive learning (HiMPC)* approach with a *hard skeleton mining (HSM)* mechanism (see Chapter 5), which exploits unlabeled hierarchical skeleton representations (*i.e.*, **coarse-to-fine body structure and motion**) of key informative skeletons to contrast and learn the most typical skeleton features for person re-ID. HiMPC mainly solves the second challenge by not only constructing **different-level** skeleton representations to enhance skeletal structure and motion learning, but also jointly learning the most typical skeleton features (defined as *"meta-prototypes"*) from multiple representation subspaces, which enhances

the robustness, consistency, and effectiveness of previous skeleton prototype contrastive learning paradigms such as the first study SimMC. Moreover, it for the first time explores the importance of each skeleton within a sequence to better learn hard and easily-confused patterns. Extensive experiments on five public benchmarks, including multi-view and RGB-based scenarios with estimated skeletons, demonstrate that our approach outperforms most state-of-the-art methods on person re-ID tasks. We further reveal the feasibility of exploiting more concise and abstract skeleton representations to perform person re-ID.

- To address the third problem to fully capture latent body relations, we present a *Transformer-based Skeleton Graph prototype contrastive learning (TranSG)* paradigm (see Chapter 6), which integrates different relational features of skeleton graphs and contrasts representative graph features to learn discriminative representations for person re-ID. To the best of our knowledge, TranSG is the first *transformer* paradigm that **unifies skeletal relation learning** and skeleton graph contrastive learning specifically for skeleton-based person re-ID, which manly addresses third challenge in this thesis. We propose the *skeleton graph transformer (SGT)* to concurrently capture structural and actional **body relations** in skeleton graphs, and devise the *graph prototype contrastive learning (GPC)* to contrast and learn representative graph features and class-related semantics at both skeleton and sequence levels, which further improves the reliability and effectiveness of prototype learning (*e.g.*, in the second study Hi-MPC) by introducing label supervision. The proposed *graph structure-trajectory prompted reconstruction (STPR)* is the first exploration of *graph prompts* in terms of structure and trajectory contexts for general skeleton semantics learning. Comprehensive experiments on five public benchmarks demonstrate the effectiveness of TranSG and its high scalability to be applied to different-level graph modeling, RGB-estimated or unlabeled skeleton data.

- To tackle the fourth challenge and achieve general model enhancement, we propose a generic **re-ranking technique** *specifically* for skeleton-based person re-ID (see Chapter 7), which encodes the most salient features of skeleton sequences based on k-reciprocal distance for **feature re-ranking** to enhance performance of different models. In the method, the *skeleton sequence pooling*

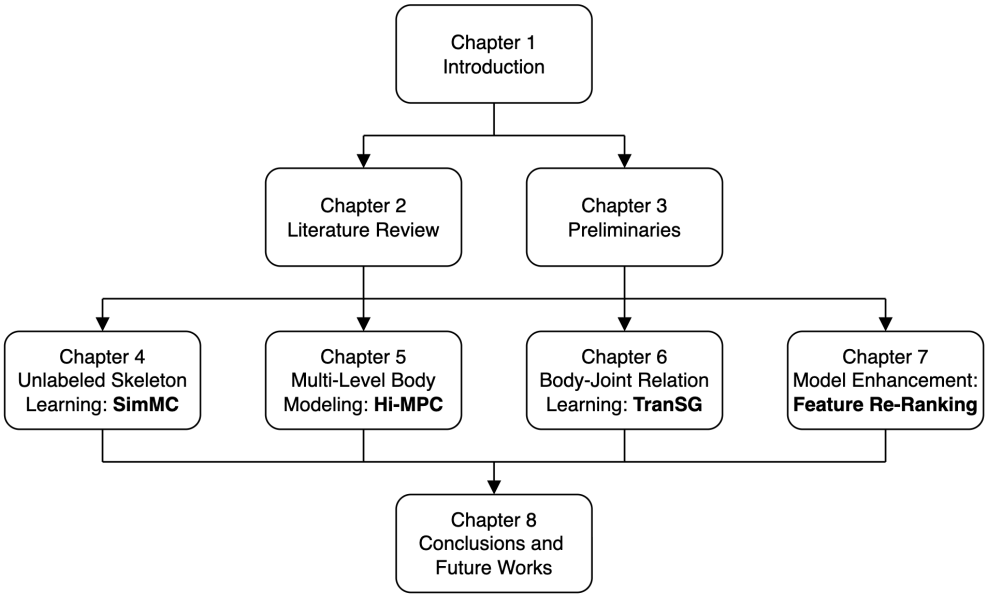


FIGURE 1.3: Overview of the hierarchical structure of the thesis.

(SSP) that combines average pooling and max pooling is proposed to aggregate the most salient features of a skeleton sequence for better ranking. It exploits the fusion of both original Euclidean distance and k-reciprocal distance, which integrates the context information of reciprocal nearest neighbors, for skeleton representation re-ranking. Moreover, we devise the *context-based Rank-1 voting* to jointly utilizes the local context (*i.e.*, top-k candidates) of gallery representations in both initial ranking list and re-ranking list to better select the Rank-1 candidate for person re-ID. Experiments on different benchmarks show that our method is highly effective on re-ranking various state-of-the-art skeleton representations to improve their performance.

1.4 Thesis Organization

An overview for the structure of the thesis is presented in Fig. 1.3, which is organized as follows:

- **Chapter 1** introduces the background of skeleton-based person re-ID, highlights the significance and motivation to develop AI-empowered skeleton-based person re-ID models, and overviews the existing methods in this area. This chapter also identifies the core research problem of the thesis, limitations

of previous methods, as well as the key challenges in skeleton-based person re-ID. Finally, we provide a brief summary of our proposed AI models and frameworks tailored for skeleton-based person re-ID under different scenarios.

- **Chapter 2** provides a literature review of existing skeleton-based person re-ID research, and summarizes the recent advances in related fields, including contrastive learning and skeleton semantics learning. It also pinpoints the shortcomings of conventional person re-ID models, and highlights the differences between the proposed approaches and existing methods.

- **Chapter 3** offers technical preliminaries for subsequent sections by introducing the models or concepts used in the proposed approaches. It presents the definitions and fundamental knowledge for the long short-term memory (LSTM), transformer, multi-level skeleton graphs, and prompt learning.

- **Chapter 4** presents a generic unsupervised framework SimMC to learn effective representations without using any label for person re-ID. We detail its two technical components, masked prototype contrastive learning (MPC) and masked intra-sequence contrastive learning (MIC), which are respectively devised for unsupervised contrastive learning of the most representative features (prototypes) within skeleton sequences, and the motion consistency contrasting between subsequences. This chapter presents comprehensive quantitative and qualitative assessments of the framework, encompassing empirical evaluations on different benchmarks, ablation studies, parameter analyses, and a theoretical analysis of two pivotal components.

- **Chapter 5** proposes an unsupervised Hi-MPC approach with a hard skeleton mining (HSM) mechanism to infer importance of skeletons and learn hierarchical coarse-to-fine skeleton representations of key informative skeletons for person re-ID. This chapter elaborates on its three key components, including hierarchical skeleton representations, hierarchical meta-prototype contrastive learning (Hi-MPC), and HSM. A thorough comparison with existing hand-crafted, supervised, self-supervised, and unsupervised methods is provided, while we conduct a systematic analysis of component effectiveness, model generality, graph representation scalability, parameter sensitivity, and theoretical interpretability in this chapter.

- **Chapter 6** presents a supervised TranSG paradigm to integrate relational features of body and motion within skeletons and learn discriminative skeleton graph representations for person re-ID. This chapter illustrates its three key modules, including skeleton graph transformer (SGT), graph prototype contrastive learning (GPC), and graph structure-trajectory prompted reconstruction (STPR), which respectively learn unified body-joint relations, discriminative graph representations, and general skeleton semantics for person re-ID. We conduct extensive experiments on five public benchmarks and compare TranSG with existing hand-crafted, graph-based, and sequence learning methods. A systematic analysis for the effects of each component and key hyper-parameters is offered, while we further verify its generality under different graph modeling, RGB-estimated skeletons, and unsupervised scenarios.

- **Chapter 7** proposes a general k-reciprocal distance based skeleton feature re-ranking method to improve the performance of skeleton-based person re-ID by optimizing the ranking of skeleton representations in the gallery. This chapter details the proposed mechanisms of skeleton sequence pooling (SSP), fused k-reciprocal distance based re-ranking, and context-based Rank-1 voting, which aim to aggregate the most salient features of a skeleton sequence and fully exploit the local context information of nearest neighbors to better re-rank skeleton features for person re-ID. We provide both empirical evaluation and qualitative analysis to demonstrate the effectiveness and generality of the proposed method when applied to existing state-of-the-art methods.

- **Chapter 8** concludes this thesis by summarizing our research contributions, future research directions, research impacts and potential applications.

Chapter 2

Literature Review

In this chapter, we first systematically review existing conventional RGB-based person re-ID methods using images/videos (Sec. 2.1.1) and skeleton-based person re-ID methods (Sec. 2.1.2). We also briefly summarize the person re-ID methods that utilize other modalities or/and their combinations (Sec. 2.1.3). Then, we present the relevant works in the field of contrastive learning (Sec. 2.2), and skeleton semantics learning (Sec. 2.3).

2.1 Person Re-Identification (Re-ID)

2.1.1 RGB-Based Person Re-ID

Person re-ID models based on RGB images and videos have been widely investigated and applied to many scenarios in the past decades [12, 41, 42]. Early works typically focus on devising discriminative hand-crafted descriptors [12, 43, 44] and/or robust distance metric learning paradigms [41, 45–47]. Recent years have witnessed the great success of deep learning based architectures in improving the person re-ID performance on widely-used image and video benchmarks such as CUHK03 [48], Market-1501 [49], and MSMT17 [50]. The mainstream RGB-based models using deep learning can be mainly grouped into supervised models and unsupervised models. The supervised RGB-based methods typically exploit local (*e.g.*, part-based) features [51], data augmentation and synthesis [52], visual attention mechanisms [53], or prior human semantics [54] to tackle the challenges of

cluttered background, occlusion, viewpoint changes, and illumination variations in image and video based scenarios. The unsupervised RGB-based methods either directly learn discriminative pedestrian representations without any external labeled data [55–57], or use domain adaptation techniques [58–60] to transfer identity-related knowledge from labeled source domain to unlabeled target domain. These methods commonly adopt K-Nearest Neighbors (KNN) [59], clustering mechanisms [55, 56], or graph-based strategies [57] to associate identities across cameras.

Gaps and Limitations. The RGB-based methods often have complex network architectures (*e.g.*, deep CNN backbones) and require huge computational resources to process the input image and video data, which possess much larger input size than the skeleton data. Another unavoidable challenge is that these methods generally require disentangling high-quality appearances, colors or textures from unrelated or noisy environmental factors (*e.g.*, cluttered background, occlusion, illumination, viewpoint) to recognize persons, while their performance could be unstable under the changes of appearances and environments.

In contrast, AI-empowered person re-ID models based on skeleton data not only have smaller network scales [5] but also are more robust to scale and view variations, as they do not require any appearance information [30]. These advantages enable them to be applied to more general scenarios with a protection of appearance-related privacy.

2.1.2 Skeleton-Based Person Re-ID

2.1.2.1 Hand-Crafted Methods

Most previous skeleton-based methods manually extract 3D skeleton features from anthropometric and gait aspects to depict human body and motion patterns [1, 2, 24, 34]. In [34], seven Euclidean distances between certain joints are integrated into a learnable distance matrix for person re-ID. They are further extended into 13 (D_{13}) and 16 skeleton descriptors (D_{16}) in [1] and [2] respectively, which are learned by different classifiers (k-nearest neighbor, support vector machine or Adaboost) to perform person re-ID tasks.

Gaps and Limitations. These methods rely heavily on domain expertise such as anatomy and kinematics [35] to model skeleton data, and lack the flexibility to

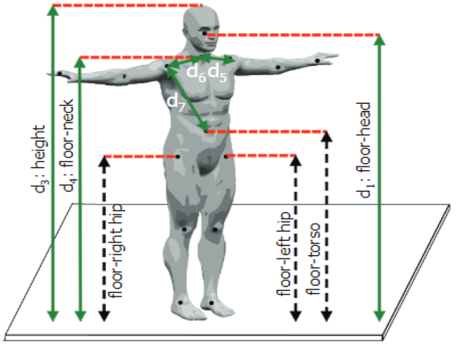


FIGURE 2.1: Geodesic distances of human body extracted for person re-ID [34].

fully exploit latent features beyond human cognition. For example, they manually extract joint distances or geodesic distances (shown in Fig. 2.1) from human body as the feature descriptor for person re-ID. Due to the limited performance of existing descriptors and their inherent requirement of domain expertise, these methods are usually combined with more efficient features such as 3D point clouds [23] or face descriptors [2] to boost accuracy.

Unlike previous hand-crafted studies, the thesis mainly focuses on devising AI models and algorithms that can perform automatic skeleton representation learning for person re-ID. The proposed methods do not require manual skeleton feature engineering but adaptively learn optimized high-dimensional representations through novel objective functions.

2.1.2.2 Deep Learning Based Methods

Recent years have witnessed the great success of deep learning in supervised and self-supervised skeleton representation learning [3, 4, 7, 25]. A self-supervised skeleton encoding model (AGE) with locality-aware attention based LSTM [3] is devised to encode discriminative gait patterns for person re-ID. The AGE model is shown in Fig. 2.2. It leverages an LSTM-based encoder-decoder architecture to perform reverse skeleton sequence reconstruction, in which the model can adaptively focus on the most important gait state to facilitate skeleton reconstruction and semantics learning (*e.g.*, discriminative patterns). However, this model is based on the assumption of intra-sequence locality and can only capture the local relations within a sequence, which limits its ability to mine valuable sequence-level features. This problem is studied by Rao *et al.* [4], and they propose an extension model

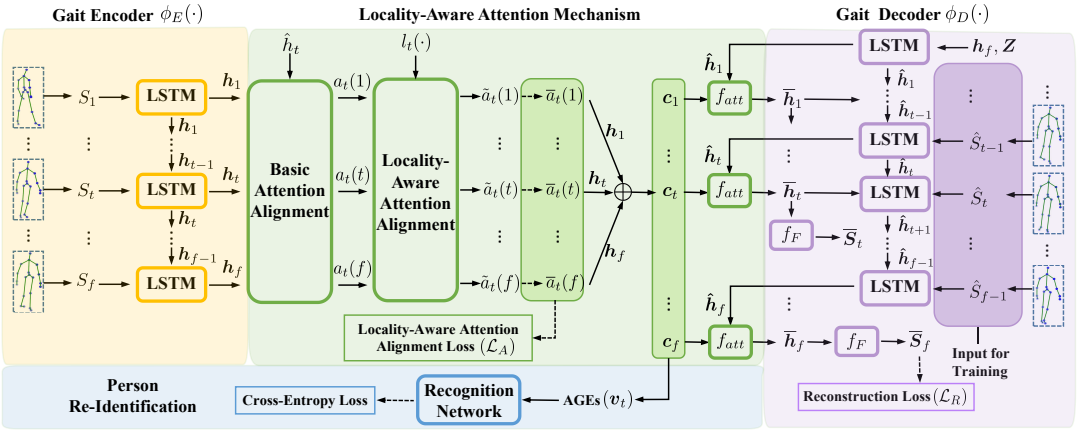


FIGURE 2.2: Overview of self-supervised Attention-based Gait Encoding (AGE) model [3] based on skeleton data.

SGELA, which combines inter-sequence contrastive learning to further enhance the relation learning between different sequences. The SGELA model also incorporates diverse pretext tasks such as sorting and prediction into gait encoding process to capture more effective skeleton features and high-level semantics for person re-ID. However, this method requires manually devising semantic pretext tasks and separately training gait encoding models for different tasks, resulting in less model generality and learning efficiency. Moreover, the inter-sequence contrastive learning only considers continuous sequences as positive pairs for similarity learning, while it lacks the ability to capture global relations (*e.g.*, class-associated feature distribution) of different sequences.

In [25], a CNN-based architecture PoseGait is leveraged to learn pre-defined skeleton/pose features for supervised human recognition. In particular, the PoseGait model extracts human body pose features (2D and 3D pose features) and spatio-temporal features (joint angles, limb lengths, joint motion), totally 81 features from skeleton sequences to learn effective gait representations. The used joint angle information is visualized on the left side of Fig. 2.3. Most of these features are manually selected based on domain knowledge as used in hand-crafted methods [1, 2, 24], thus PoseGait can be viewed as a semi-AI model that combines both prior knowledge and automatic representation learning architectures.

In [7], skeleton graphs are constructed based on the physical connections of body joints or parts to learn skeletal relations (*e.g.*, structural and collaborative relations) and high-level motion semantics for person re-ID. The SM-SGE framework [8] further integrates multi-scale skeleton reconstruction and cross-scale skeleton

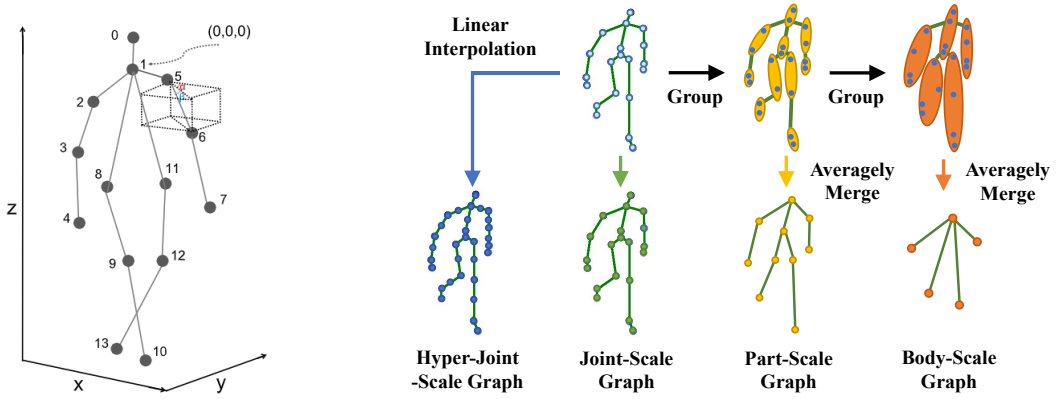


FIGURE 2.3: Visual examples of body-joint angle features used in the PoseGait model [25] (Left side) and multi-scale skeleton graph construction used in the SM-SGE framework [8] (Right side).

inference into a graph encoding framework for self-supervised person re-ID. As shown on the right side of Fig. 2.3, this framework not only designs coarse-to-fine skeleton graphs (joint-scale, part-scale, and body-scale graphs) based on the original skeleton, but also explores denser skeleton representations (hyper-joint-scale graph) by applying linear interpolation. Compared with single-level skeleton representations, the used multi-scale skeleton graphs is shown to benefit the skeleton feature and relation learning for person re-ID tasks in [8]. In [61], an unsupervised sequence-level skeleton prototype learning approach (SPC-MGR) that alternates feature clustering and contrasting is incorporated with multi-level relation learning [7] for person re-ID.

Comparison with Existing Methods. Unlike previous hand-crafted, supervised, and self-supervised models, the proposed models SimMC (see Chapter 4) and Hi-MPC (see Chapter 5) require no pre-defined skeleton descriptors or pretext design for skeleton representation learning. They can efficiently exploit *unlabeled* 3D skeleton data to mine the most representative features and contrast their inherent relationships for person re-ID. To improve these models by generating skeleton prototypes (*i.e.*, representative skeleton features/clusters) more reliably, we further propose TranSG (see Chapter 6) that exploits the graph feature centroid of each *ground-truth* identity as prototypes, which are utilized to guide the skeleton contrastive learning in a supervised manner for for person re-ID.

Comparison with Methods Combining Skeleton Data. The skeleton/pose data are also utilized or combined in many image/video-based person re-ID methods to help extract pre-defined local/hierarchical body parts [62, 63], disentangle

semantic components (*e.g.*, bone locations) [64, 65] or generate pose augmented representations [66]. The key differences between these methods and our approaches are two-fold. First, they use skeleton data as *auxiliary* information to combine with RGB images to boost the model performance, while our work exploits *only* skeleton data without using any appearance-based features for person re-ID. Second, they typically leverage the hierarchical structure of the original body skeletons to generate image-based body parts, while our approaches such as Hi-MPC (see Chapter 5) and TranSG (see Chapter 6) can hierarchically or individually construct skeleton representations at different levels, each of which corresponds to a new *independent* body representation to learn discriminative pattern information for person re-ID.

2.1.3 Person Re-ID Using Other Modalities

Some existing works utilize other modalities such as depth images and 3D point clouds or their combination (*i.e.*, multi-modal data) to extract human body shapes, silhouettes or motion trajectories for person re-ID. Sivapala *et al.* [67] extend the Gait Energy Image (GEI) [68] to 3D domain and proposes Gait Energy Volume (GEV) algorithm based on depth images to perform gait-based human recognition. Munaro *et al.* [1] propose point cloud matching (PCM) to compute the distances of multi-view point cloud sets, so as to discriminate different persons. Haque *et al.* [69] adopt 3D LSTM to model motion dynamics of 3D point clouds for person Re-ID. As to multi-modal methods, they usually combine skeleton-based features with extra RGB or depth information (*e.g.*, depth shape features based on point clouds [23, 70, 71]) to boost Re-ID performance. For example, some works combine RGB images and skeleton data to learn auxiliary anthropometric attributes [72], body parts correlations [65], and clothing-invariant features [73] to enhance their performance. In [74], CNN-LSTM with reinforced temporal attention (RTA) is proposed for person Re-ID based on a split-rate RGB-depth transfer approach. There also exist some methods that re-identify persons across two different modalities or scenarios, including text-to-image [75, 76], visible-to-infrared [77, 78] and cross-resolution person re-ID [79, 80].

2.2 Contrastive Learning

The aim of contrastive learning is to pull closer homogeneous or positive representation pairs while pushing farther negative pairs in a certain feature space. It has been broadly applied to various areas to learn effective feature representations [29, 81–84]. Mainstream contrastive learning paradigms can be mainly grouped into two categories, instance-wise contrastive learning [4, 81, 85, 86] and prototypical contrastive learning [29, 87–89]. We briefly review related works on these two types of paradigms, and also offer an overview of hard negative mining mechanisms designed for contrastive learning paradigms.

2.2.1 Instance-Wise Contrastive Learning

In the instance-wise contrastive learning (ICL), the model brings together representations of different views (*e.g.*, augmented samples) from the the same instance, while pushing representations of different instances apart using instance-level contrastive losses. As a representative form of ICL, the instance discrimination method with exemplar tasks and noise-contrastive estimation (NCE) [90] is proposed in [81] for visual contrastive learning. The common practice in ICL is to utilize a large batch size to generate positive and negative instance pairs [91, 92], or a queue-based dictionary on the fly to update negative instances using momentum encoders [93, 94]. A Siamese architecture is proposed in [95] to perform ICL based on a single positive pair without using negative pairs or momentum encoders. A locality-aware contrastive learning mechanism with consecutive skeleton sequences as instances is devised in [4] for person re-ID. Chen *et al.* [85] and Wang *et al.* [86] further explore the generalized ICL loss by enhancing the feature alignment of positive pairs and the uniformity of normalized representations on the hyper-sphere.

2.2.2 Prototypical Contrastive Learning

A few studies focus on the contrastive learning with feature prototypes assigned by clustering or pre-defined memory mechanisms. In [87], an online algorithm SwAV is devised to enforce the consistency between cluster assignments of different augmented views while predicting the code of a view from another-view representations

for image classification. The momentum-based contrastive learning and k-means clustering are combined in the PCL framework [88] to perform unsupervised visual learning. CLD [89] integrates between-instance similarity with the cross-level discrimination between instances and local instance groups to achieve better invariant mapping and contrastive learning. An offline prototype generation approach (SPCL) with a hybrid memory mechanism is proposed in [29] for domain adaptive objective re-ID tasks. Fundamentally different from previous studies that leverage augmented samples of images as contrastive instances, we devise a new generic meta-prototype contrastive learning paradigm for 3D skeleton data in Hi-MPC (see Chapter 5), which exploits unlabeled hierarchical skeleton representations as instances to mine the most representative and discriminative features (*i.e.*, skeleton meta-prototypes) for meta-prototype-instance contrastive learning.

2.2.3 Hard Negative Sample Mining and Contrasting

Hard negative sample mining aims to find more informative training samples that are difficult to discriminate (*e.g.*, easily-confused negatives), which has been widely applied to various areas to accelerate network training and improve model performance [39, 96, 97]. In contrastive learning, existing hard negative mining models can be mainly grouped into (1) *adversarial learning based methods* [98, 99] that generate hard negatives by adversarial optimization and (2) *mixing based methods* [100–102] that mix the negative samples and positive samples in the feature space. In [98], a representation network and its negative adversary are alternately trained to generate the hardest negative samples for contrasting, while Wang *et al.* [99] further introduce a diversity loss to generate diverse challenging negative samples with different noise. A hard negative mixing strategy is proposed in [100] to mix features of hardest negatives and its query to enhance contrastive learning. To improve the difficulty of generated negatives, a diversity objective function is devised in [102] to mix multiple samples with dynamic weights. Some recent works also explore controllable hard negative mining with an importance sampling strategy [96] or dynamic curriculum learning [97].

Previous methods often require extra triplet constraint [39, 103], adversarial learning [98, 99] or sample mixing [100–102]. The proposed hard skeleton mining mechanism in Hi-MPC (see Chapter 5) can directly exploit the inherent similarity between skeleton representations and cluster-level representations to adaptively infer the informative importance of skeletons at different skeleton levels and different feature subspaces, and mine both hard positive and negative skeleton samples from sequences without using any labels.

2.3 Skeleton Semantics Learning

Learning general effective and discriminative skeleton semantics is pivotal to skeleton-based person re-ID [3, 4, 7, 14] and other skeleton-related tasks (*e.g.*, 3D action recognition) [40, 104–107]. The attention-based reconstruction [3] and attention-based contrastive learning [4] are devised to help learn semantics of motion continuity within skeletons. In [7], the multi-level skeleton sequence prediction (MSSP) task is proposed to sequentially predict skeletons from multi-level graphs, while Rao *et al.* [8] further devise the multi-scale skeleton reconstruction (MSR) to capture skeleton dynamics and cross-scale component correspondence. The masked intra-sequence contrastive learning (MIC) is devised in [5] to learn motion continuity and pattern invariance from different skeleton subsequences. In our proposed model TranSG (see Chapter 6), a structure-trajectory prompted reconstruction (STPR) task is employed to learn structural relations and pattern continuity of body joints.

Existing skeleton semantics learning tasks usually rely on certain model architectures or representations [3, 4, 7, 8] and can only perform sequence-level learning without explicit spatial [3, 4] or temporal modeling [5]. In contrast, the proposed STPR (see Fig. 6.2) is able to simultaneously model 3D skeletons in terms of spatial (*e.g.*, body structural positions) and temporal features (*e.g.*, joint trajectory), which could possess higher generality to be applied to different models.

Chapter 3

Preliminaries

3.1 Long Short-Term Memory

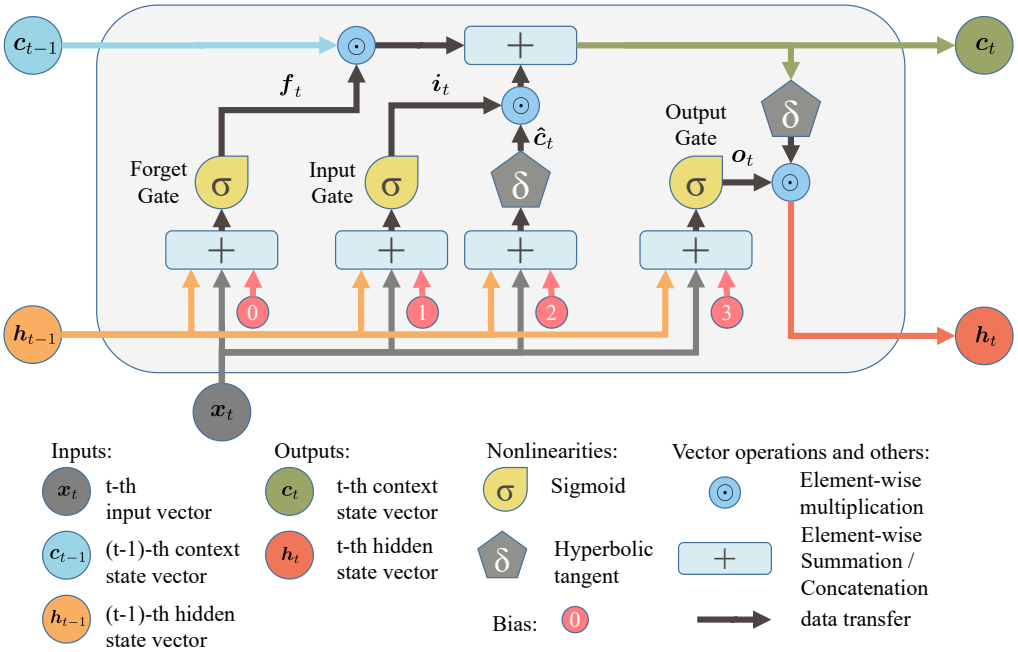


FIGURE 3.1: Schematic diagram of Long Short-Term Memory (LSTM). The explanation for symbols and operations of LSTM are provided at the bottom.

Long Short-Term Memory (LSTM) [36] has been widely utilized to capture long-term dependencies and dynamics in time series data such as 3D skeleton sequences [3, 4, 7, 69]. In this thesis, we implement and compare different existing LSTM-based models to validate the effectiveness of the proposed approaches. LSTM is a special type of Recurrent Neural Network (RNN) variant designed to tackle the

common challenge of standard RNNs, such as the vanishing gradient problem. The core idea of LSTM is the cell state, which acts like a conveyor belt transferring relevant information through the sequence processing, and three types of gates (Forget Gate, Input Gate, Output Gate) that regulate the flow of information. During the forward pass, the LSTM processes input data sequentially, updating its internal state at each timestep based on the current input and the previous output. The state updates are controlled by the gates, which are themselves updated based on the input data and the recurrent connections. Based on this mechanism, the LSTM can handle long-term dependencies by maintaining a cell state over time. The gates of the LSTM modulate the cell state by selectively adding or removing information, which helps the network to retain important information over long sequences without the gradient either vanishing or exploding.

To better illustrate the structure and function of LSTM, we provide a visualization for the LSTM unit (see Fig. 3.1), which is composed of three gates (input gate, forget gate, output gate) and two states (hidden state and context state (also called memory state)): (1) The input gate allows the input and previous hidden state to alter the candidate context state and retain the crucial input information. (2) The forget gate enables the previous context state to keep the most useful context information by selectively forgetting the state information. (3) The output gate can integrate the current context state to generate the current hidden state and adjust its effect of on the rest of network. (4) The context state and hidden state are used to store the crucial temporal information for the next encoding or output. Formally, given an input vector $\boldsymbol{x}_t$ at t^{th} time step, the LSTM encodes $\boldsymbol{x}_t$ and the previous step's hidden state $\boldsymbol{h}_{t-1}$ (note that at the first time step we use an all-zero placeholder $\boldsymbol{h}_0 = Z \in \mathbb{R}^K$) into the current hidden state $\boldsymbol{h}_t$, with the internal operations of different gates and states as below:

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}_i[\boldsymbol{x}_t; \boldsymbol{h}_{t-1}] + b_i\right) \tag{3.1}$$

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_f[\boldsymbol{x}_t; \boldsymbol{h}_{t-1}] + b_f\right) \tag{3.2}$$

$$\hat{\boldsymbol{c}}_t = \delta\left(\boldsymbol{W}_c\ [\boldsymbol{x}_t; \boldsymbol{h}_{t-1}] + b_c\right) \tag{3.3}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \hat{\boldsymbol{c}}_t \tag{3.4}$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}_o[\boldsymbol{x}_t; h_{t-1}] + b_o\right) \tag{3.5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \delta\left(\boldsymbol{c}_t\right) \tag{3.6}$$

where $\boldsymbol{x}_t, \boldsymbol{h}_t, \hat{\boldsymbol{c}}_t, \boldsymbol{c}_t$ are the input vector, hidden state vector, candidate context state vector, and context state vector at time t respectively. $\boldsymbol{i}_t, \boldsymbol{f}_t, \boldsymbol{o}_t$ denote activation vectors output from the input gate, forget gate, and output gate at time t. $\boldsymbol{W}_i, \boldsymbol{W}_f, \boldsymbol{W}_c, \boldsymbol{W}_o$ are weight matrices to be learned. b_i, b_f, b_c, b_o are bia vectors. $\sigma(\cdot)$ is the sigmoid function and δ is the tanh function. $\odot$ represents the scalar product of two vectors. In our work, the embedding size of $\boldsymbol{h}_t$, $\hat{\boldsymbol{c}}_t$, and $\boldsymbol{c}_t$ is set to the same K. It should be noted that for convenience and clarity of presentation, we adopt the commonly used denotations (*e.g.*, $\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{c}$) in Eqs. 3.1-3.6, which are NOT the formal definition used in other chapters.

3.2 Transformer

Transformer [108] is a widely-used paradigm in various fields of natural language processing (NLP), computer vision (CV), speech recognition, *etc* [109–112]. In this thesis, we explore a new transformer paradigm that unifies structural and actional relation learning specifically for skeleton-based person re-ID in Chapter 6. The standard transformer leverages self-attention mechanisms to effectively capture complex dependencies in data, making it powerful for tasks involving large amounts of sequential data. We present the overview of a transformer paradigm in Fig. 3.2. The process of transformer learning can be summarized as:

(1) *Input Representations of Tokens*: Each token in the input sequence, *e.g.*, words in a sentence, is converted into a numerical representation (termed as embedding), like $\boldsymbol{h}_t$ shown in Fig. 3.2. These embeddings are learnable during training to achieve the model objective.

(2) *Positional Encoding*: The transformer does not inherently process data sequentially but parallelly, thus they require a mechanism to understand the order of tokens. To this end, positional encodings are added to the input token embeddings to inject context information about the position of each token in the sequence. This helps the model to maintain awareness of the sequence order. Note that this step is not required for some non-sequential tasks (*e.g.*, static image classification), and we do not show it in the paradigm for simplicity.

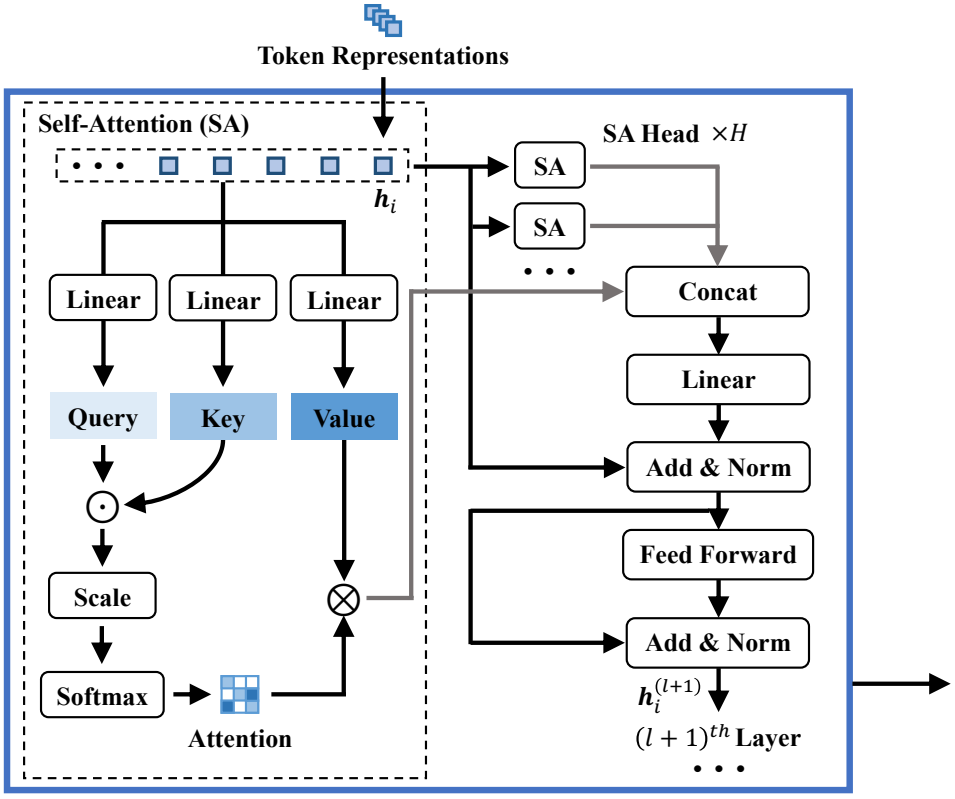


FIGURE 3.2: Schematic diagram of transformer. "Scale": Scaled dot-product similarity computation; "Concat": Concatenation of features from different SA heads; "Linear": Linear transformation; "Feed Forward": Feed Forward Network (FFN) with residual connections and batch normalization. $\boldsymbol{h}_i^{l+1}$ denotes the i^{th} token representation in the $(l+1)^{th}$ layer, and H represents the number of SA heads.

(3) *Self-Attention Mechanism*: The self-attention mechanism allows the model to weigh the importance of all other tokens in the sequence for each token. For each token, the transformer calculates attention scores relative to every other token, and determines which tokens are most relevant during processing. This is crucial for tasks like understanding context in sentences or dependencies in data. The computation of self-attention is often integrated into a single self-attention (SA) head, and the transformer generally adopts multi-head attention mechanisms (*e.g.*, H heads in Fig. 3.2) to concatenate features independently learned from different SA heads.

(4) *Layered Architecture*: The transformer learns representations through multiple layers of multi-head attention and feedforward neural networks. Each layer operates on the entire sequence simultaneously, enabling parallel processing and faster computation times compared to LSTM (see Fig. 3.1). The outputs from each

layer are sequentially passed to the next, which allows the model to build complex representations.

(5) *Output Sequence*: After processing through multiple layers, the transformer generates an output sequence that can be used for various downstream tasks like translation, summarization, or classification. It is worth noting that each output token is influenced by all input tokens, adjusted according to the learned attention weights.

3.3 Multi-Level Skeleton Graphs

Human body can be generally segmented into numerous functional components with diverse granularities (*e.g.*, knee joint, thigh part, leg limb), each of which carries different geometric or anthropometric attributes of body [7, 113, 114]. Inspired by this fact, many works [7, 8, 61] regard body joints as the basic components, and merge spatially nearby groups of joints to be a higher level body-component node at the center of their positions.

As shown in Fig. 3.3, a common practice is to construct multi-level skeleton graphs at three levels, namely *joint-level*, *part-level*, and *body-level* graphs for each skeleton [7, 8]. More specifically shown in Fig. 3.4, for a skeleton with 20 body joints, we pre-define the indices for each joint, and divide them into different parts based on manual rules or domain knowledge. We averagely merge the positions of all joints within a particular part to construct a new node, thereby building part-level graphs (10 nodes) and body-level graphs (5 nodes). For joint-level graphs (20 nodes), we keep each joint as a node to construct graphs.

Formally, the construction of multi-level skeleton graphs can be formulated as follows: Each level graph $\mathcal{G}^m(\mathcal{V}^m, \mathcal{E}^m)$ $(m \in \{0, 1, 2\})$ consists of nodes $\mathcal{V}^m = \{\boldsymbol{v}_1^m, \boldsymbol{v}_2^m, \cdots, \boldsymbol{v}_{n_m}^m\}$ $(\boldsymbol{v}_i^m \in \mathbb{R}^D,\ i \in \{1, \cdots, n_m\})$ and edges $\mathcal{E}^m = \{e_{i,j}^m \mid \boldsymbol{v}_i^m, \boldsymbol{v}_j^m \in \mathcal{V}^m\}$ $(e_{i,j}^m \in \mathbb{R})$. Here m denotes the index of graph level, and $0, 1, 2$ respectively indicate the original joint-level, part-level, and body-level graphs. $\mathcal{V}^m$, $\mathcal{E}^m$ denote the set of nodes corresponding to different body components and the set of their relations respectively, and n_m is the number of nodes in the m^{th} level graph $\mathcal{G}^m$. We use $\mathbf{A}^m \in \mathbb{R}^{n_m \times n_m}$ to represent a graph's adjacency matrix, where each element $\mathbf{A}_{i,j}^m$ is defined as the relational value between nodes i and j. $\mathbf{A}^m$ is initialized

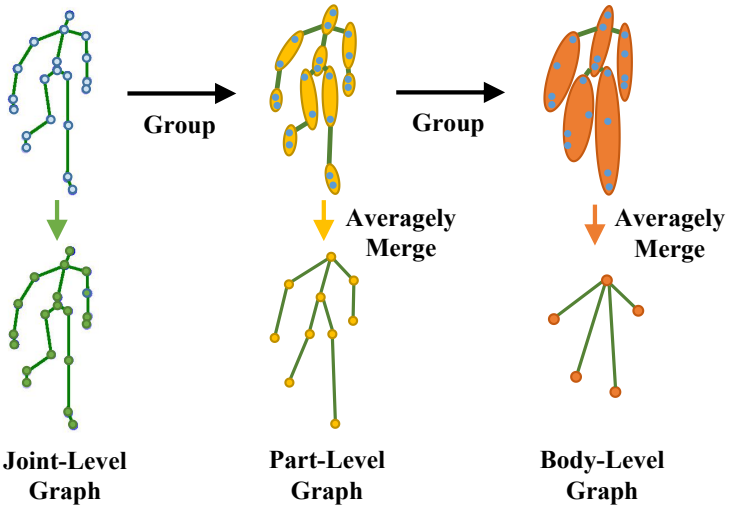


FIGURE 3.3: Examples of three graph scales for a skeleton with 20 body joints. We divide body into 10 and 5 parts to build part-level and body-level graphs, and merge internal joints into graph nodes [7, 8].

based on the connections of adjacent body joints, which is *learnable* during model training to capture dynamic relations and valuable skeleton patterns.

It worth noting that body partitions in the above three-level skeleton graphs are pre-defined based on human prior knowledge and previous common practice, which theoretically can be extended to arbitrary number of levels and diverse topological structures [114]. The low-level skeleton representations such as joint-level graphs contain detailed *local* body structure and positional information, while the high-level skeleton representations such as body-level graphs can be viewed as to further abstract the human body to focus on more *global* features of four body limbs and torso. Therefore, the multi-level skeleton graphs essentially help models to characterize *coarse-to-fine* (*e.g.*, global-to-local) body and motion features to jointly learn more discriminative person re-ID representations. In this thesis, we construct different-level skeleton graphs in Chapters 5 and 6 to evaluate the effectiveness and generality of Hi-MPC and TranSG methods when applied to skeleton graphs with varying scales.

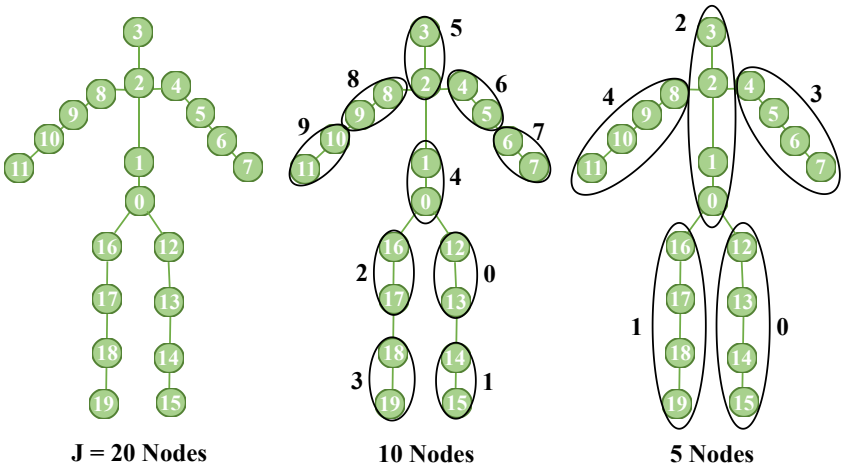


FIGURE 3.4: Node indices for joint-level (20 nodes), part-level (10 nodes), and body-level (5 nodes) graph representations of skeletons from IAS, BIWI and KGBD datasets. The indices correspond to the following joint names: (0) "SpineBase", (1) "SpineMid", (2) "Neck", (3) "Head", (4) "LeftShoulder", (5) "LeftElbow", (6) "LeftWrist", (7) "LeftHand", (8) "RightShoulder", (9) "RightElbow", (10) "RightWrist", (11) "RightHand", (12) "LeftHip", (13) "LeftKnee", (14) "LeftAnkle", (15) "LeftFoot", (16) "RightHip", (17) "RightKnee", (18) "RightAnkle", (19) "RightFoot".

3.4 Prompt Learning

Prompt learning is originally defined by [115] as "fill-in-the-blank" cloze tests for knowledge probing, which has been widely applied to various areas including vision-language representation learning [83, 84, 116]. The function of prompts is to provide additional knowledge, instruction, or context for the input of models, such that they can be prompted to give more reliable outputs for different tasks [84, 115, 117, 118]. In [83], CLIP leverages language-based prompts to generalize the pre-trained visual representations to many tasks. CoOp [116] is further devised to automatically model task-relevant prompts with continuous representations to improve the downstream task performance. As far as we know, the proposed TranSG (see Chapter 6) for the first time explores *graph prompts* (defined as structure and trajectory contexts) for skeleton graph reconstruction, so as to encourage capturing more key features and graph semantics (*e.g.*, pattern continuity) for person re-ID.

Chapter 4

Skeleton-Based Person Re-ID with Unlabeled Skeleton Learning

4.1 Introduction

In the past few years, person re-ID via 3D skeletons has drawn growing interests from both academia and industry [2, 23–25]. Despite the progress in this area, existing endeavors require either extracting hand-crafted features (*e.g.*, anthropometric attributes) [2, 24, 34] or learning skeleton representations with the supervision of labels [3, 7, 25]. These methods usually require a specific pre-modeling of 3D skeletons (*e.g.*, skeleton graphs [7]), and rely on massive manually-annotated data to train or fine-tune models, which is labor-expensive and unable to learn general pedestrian representations under the unavailability of labels. In other words, these previous methods lack the ability to perform automatic *unsupervised* skeleton representation learning with AI architectures (corresponding to the first challenge in Sec 1.2).

To address these challenges, we present a generic Simple Masked Contrastive learning (SimMC) framework with masked prototype and intra-sequence contrastive

This chapter has been published as: Haocong Rao and Chunyan Miao, "SimMC: Simple Masked Contrastive Learning of Skeleton Representations for Unsupervised Person Re-Identification," In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 2022 [5]. DOI: 10.24963/ijcai.2022/180.

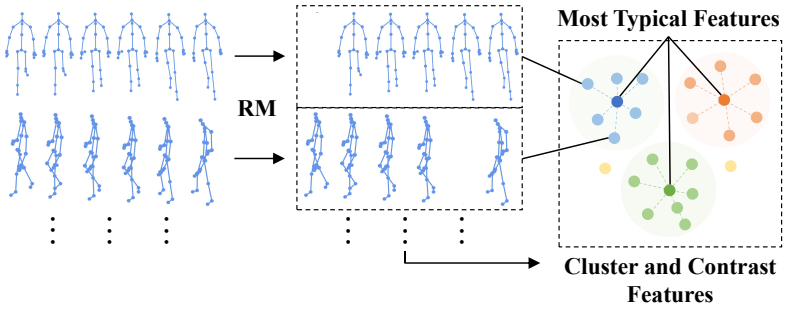


FIGURE 4.1: Simplified process of our approach: It clusters the randomly masked (RM) skeleton sequences, and contrasts their features with the most typical ones to learn discriminative skeleton representations for person re-ID. The overview of SimMC framework is provided in Fig. 4.2.

learning[1]. As shown in Fig. 4.1, SimMC contrasts the typical features and inherent relationships of *randomly-masked* (RM) skeleton sequences to learn effective skeleton representations *without using any label* for person re-ID. Specifically, to fully utilize unique features within skeleton sequences, we first devise a ***masked prototype contrastive learning (MPC)*** scheme to cluster *subsequence* representations (referred as *skeleton instances*) randomly masked from raw sequences, and contrast the inherent similarity between them and the most typical features (referred as *skeleton prototypes*) to learn discriminative skeleton representations. By pulling closer skeleton instances belonging to the same prototype and pushing apart instances of different prototypes with the instance-prototype contrastive learning, MPC enables the model to capture discriminative skeleton features and high-level semantics (*e.g.*, intra-class skeleton similarity) from *unlabeled* skeleton sequences for the person re-ID task. Then, motivated by the nature of motion continuity that typically endows different subsequences with strong correlations (*e.g.*, motion similarity), we propose the ***masked intra-sequence contrastive learning (MIC)*** to learn the intra-sequence similarity between subsequences of the same skeleton sequence, which encourages capturing the pattern consistency within sequences to learn more effective representations of skeletons for person re-ID. Empirical evaluations show that SimMC significantly outperforms most state-of-the-art skeleton-based methods on four benchmark datasets, and can be exploited to fine-tune existing skeleton representations and boost their performance with up to 28.2% mAP gains.

[1]Our codes are publicly available at https://github.com/Kali-Hac/SimMC.

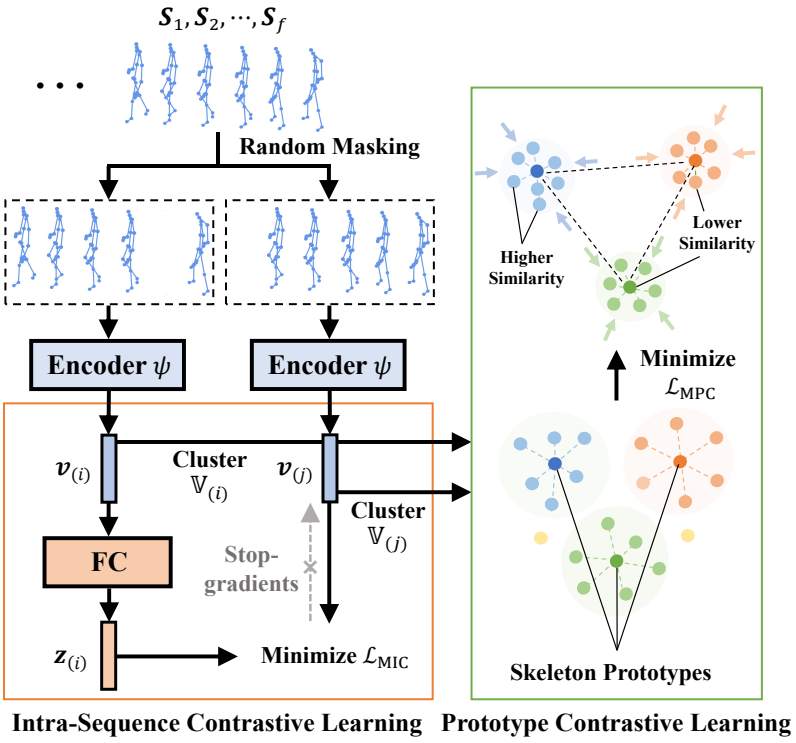


FIGURE 4.2: Schematic diagram of our framework with prototype contrastive learning and intra-sequence contrastive learning (described in Sec. 4.2.2).

Advantages. The proposed SimMC framework enjoys merits in terms of architectures, performance, and scalability. Firstly, SimMC is primarily built by multi-layer perceptron (MLP) networks with small model complexity, which can directly learn effective representations from raw skeleton sequences without any prior modeling. Secondly, the proposed unsupervised framework outperforms most existing self-supervised and supervised skeleton-based methods that utilize extra label information, and can also be efficiently applied to 3D skeleton data estimated from RGB-based scenes. Lastly, our framework can serve as a generic contrastive learning paradigm to fine-tune skeleton features learned from existing models, which benefits learning better skeleton representations for the task of person re-ID.

With this chapter, we make the following contributions:

- We present a simple masked contrastive learning (SimMC) framework that exploits typical features and relationships of masked unlabeled skeleton sequences to learn discriminative representations for person re-ID.

- We devise a novel masked prototype contrastive learning (MPC) scheme to fully contrast most representative features and learn high-level semantics from subsequence representations masked from skeleton sequences.
- We propose the masked intra-sequence contrastive learning (MIC) to learn inherent similarity and pattern consistency between subsequences, so as to encourage learning more effective representations for person re-ID.
- Empirical evaluations show that SimMC significantly outperforms most state-of-the-art skeleton-based methods on four benchmark datasets, and can be exploited to fine-tune existing skeleton representations and boost their performance with up to 28.2% mAP gains.

4.2 The Proposed SimMC Framework

4.2.1 Problem Definition

Suppose that a 3D skeleton sequence $\boldsymbol{S}_{1:f}=(\boldsymbol{S}_1,\cdots,\boldsymbol{S}_f)\in\mathbb{R}^{f\times K}$, where $\boldsymbol{S}_t\in\mathbb{R}^K$ is the t^{th} skeleton with 3D coordinates of J body joints and $K=J\times 3$. Each skeleton sequence $\boldsymbol{S}_{1:f}$ belongs to an identity y, where y $\in\{1,\cdots,I\}$ and I is the number of different identities. The training set $\Phi_{\mathcal{T}}=\left\{\boldsymbol{S}_{1:f}^{\mathcal{T},i}\right\}_{i=1}^{N_1}$, probe set $\Phi_{\mathcal{P}}=\left\{\boldsymbol{S}_{1:f}^{\mathcal{P},i}\right\}_{i=1}^{N_2}$, and gallery set $\Phi_{\mathcal{G}}=\left\{\boldsymbol{S}_{1:f}^{\mathcal{G},i}\right\}_{i=1}^{N_3}$ contain N_1, N_2, and N_3 skeleton sequences of different persons in different views and scenes. Our framework aims at learning an encoder (denoted as $\psi(\cdot)$) built with neural networks to encode $\Phi_{\mathcal{P}}$ and $\Phi_{\mathcal{G}}$ into effective skeleton representations $\{\boldsymbol{v}_i^{\mathcal{P}}\}_{i=1}^{N_2}$ and $\{\boldsymbol{v}_j^{\mathcal{G}}\}_{j=1}^{N_3}$, such that the representation $\boldsymbol{v}_i^{\mathcal{P}}$ in probe set can match the representation $\boldsymbol{v}_j^{\mathcal{G}}$ of the same identity in gallery set. The overview of our framework is presented in Fig. 4.2.

4.2.2 Overview of SimMC

The schematic diagram of SimMC is shown in Fig. 4.2: Firstly, in the part of skeleton subsequence generation, we randomly mask each input skeleton sequence to sample i^{th} and j^{th} subsequences. These skeleton subsequences are then encoded

into skeleton instances $\boldsymbol{v}_{(i)}$ and $\boldsymbol{v}_{(j)}$ (see Sec. 4.2.3). Secondly, in the part of prototype contrastive learning, we cluster corresponding instance sets $\mathbb{V}_{(i)}$ and $\mathbb{V}_{(j)}$ individually to generate skeleton prototypes, and then enhance the similarity between instances of same prototype while maximizing the dissimilarity between different ones by minimizing the proposed masked prototype contrastive learning loss $\mathcal{L}_{\text{MPC}}$ (detailed in Sec. 4.2.3). Meanwhile, in the part of intra-sequence contrastive learning, a Siamese architecture is exploited to learn inherent intra-sequence similarity between $\boldsymbol{v}_{(i)}$ and $\boldsymbol{v}_{(j)}$ by minimizing the proposed intra-sequence contrastive learning loss $\mathcal{L}_{\text{MIC}}$ (detailed in Sec. 4.2.4).

4.2.3 Masked Prototype Contrastive Learning

Each person's skeletons typically possess unique features (*e.g.*, anthropometric attributes), while their corresponding sequences could carry recognizable and highly consistent walking patterns [37]. Naturally, we expect the model to exploit the most representative skeleton patterns and traits *within each sequence* for person re-ID. A naïve solution is to cluster skeleton sequences to learn the representative features by direct inter-sequence contrastive learning, while it could overlook some valuable *intra-sequence* representations (*e.g.*, subsequences) that might contain key patterns. To encourage the model to fully mine intra-sequence skeleton features and high-level semantics (*e.g.*, identity-related patterns) from skeleton sequences, we propose a ***masked prototype contrastive learning (MPC) scheme*** to *jointly* focus on the most typical features (***skeleton prototypes***) of different subsequence representations (***skeleton instances***) randomly masked from original sequences, and exploit the instance-prototype similarity and dissimilarity to learn discriminative skeleton representations.

Given an input skeleton sequence $\boldsymbol{S}_{1:f} = (\boldsymbol{S}_1, \cdots, \boldsymbol{S}_f)$, we exploit an MLP encoder with one hidden layer to encode each skeleton as:

$$\boldsymbol{h}_j = \psi\left(\boldsymbol{S}_j\right) = \mathbf{W}^2\sigma\left(\mathbf{W}^1\boldsymbol{S}_j\right), \tag{4.1}$$

where $\psi(\cdot)$ represents the encoder function, $\mathbf{W}^1 \in \mathbb{R}^{H\times K}$ and $\mathbf{W}^2 \in \mathbb{R}^{H\times H}$ denote the learnable weight matrices to encode the j^{th} skeleton $\boldsymbol{S}_j \in \mathbb{R}^K$ into a latent feature representation $\boldsymbol{h}_j \in \mathbb{R}^H$, and $\sigma(\cdot)$ is a ReLU non-linear activation function.

Then, to sample subsequence representations from the encoded sequence representation $(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_f)$ of $\boldsymbol{S}_{1:f}$, we utilize a masking function $\mathcal{M}$ to randomly produce x masks, *i.e.*, zero-masking positions, for each skeleton sequence of length f with:

$$\mathcal{M}(f, x) = (m_1, \cdots, m_f), \tag{4.2}$$

where $m_j \in \{0, 1\}$ is the mask status for the j^{th} position of a sequence and $\sum_{j=1}^{f} m_j = f - x$. We apply the generated random masks to $\boldsymbol{S}_{1:f}$ and its corresponding skeleton representations $(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_f)$ (see Eq. (4.1)), which are then integrated into a subsequence representation as (see Fig. 4.2):

$$\boldsymbol{v}_{(i)} = \frac{1}{f - x} \sum_{j=1}^{f} m_{(i),j} w_j \boldsymbol{h}_j, \tag{4.3}$$

where $\boldsymbol{v}_{(i)} \in \mathbb{R}^H$ $(i \in \{1, \cdots, q\})$ denotes the feature representation of i^{th} subsequence sampled from $\boldsymbol{S}_{1:f}$ using x random masks, q is the number of subsequence sampling, $m_{(i),j}$ denotes the mask status of the j^{th} position at the i^{th} sampling, while w_j represents the importance of j^{th} skeleton representation $\boldsymbol{h}_j$. Here each skeleton is assumed to equally contribute to representing sequence features, *i.e.*, $w_j = 1$. For clarity, we use $\mathbb{V}_{(i)} = \{\boldsymbol{v}_{(i),j}\}_{j=1}^{N_1}$ to denote all subsequence representations in the i^{th} subsequence sampling of the training set $\Phi_{\mathcal{T}}$. Note that we sample one random subsequence for each training sequence at each sampling. $\mathbb{V}_{(i)} = \{\boldsymbol{v}_{(i),j}\}_{j=1}^{N_1}$ are exploited as *skeleton instances* for the MPC scheme.

To group feature-similar skeleton instances and discover semantic clusters with arbitrary shapes, we leverage the DBSCAN algorithm [6] to perform clustering *individually* on the i^{th} instance set $\mathbb{V}_{(i)}$ corresponding to i^{th} subsequence sampling, as shown in Fig. 4.2, and generate clusters $\overline{\mathbb{V}}_{(i)}^{c} = \{\boldsymbol{v}_{(i),j}^{c}\}_{j=1}^{N_c}$, $c \in \{1, \cdots, C\}$, where C is the number of clusters (*i.e.*, pseudo classes), and each cluster $\overline{\mathbb{V}}_{(i)}^{c}$ contains N_c instances belonging to the c^{th} pseudo class. We *averagely aggregate* instance features of the same cluster to generate the corresponding skeleton prototype as:

$$\boldsymbol{p}_{(i)}^{c} = \frac{1}{N_c} \sum_{j=1}^{N_c} \boldsymbol{v}_{(i),j}^{c} \ , \tag{4.4}$$

where $\boldsymbol{p}_{(i)}^{c} \in \mathbb{R}^H$ denotes the skeleton prototype of the c^{th} cluster $\overline{\mathbb{V}}_{(i)}^{c}$. To jointly

focus on the representative skeleton features in all instance sets and encourage capturing high-level skeleton semantics from different prototypes, we exploit a masked prototype contrastive (MPC) loss to enhance the similarity of each skeleton instance to the corresponding prototype and maximize its dissimilarity to other prototypes by:

$$\mathcal{L}_{\text{MPC}} = \frac{1}{N} \sum_{i=1}^{q} \sum_{c=1}^{C_i} \sum_{j=1}^{N_c} -\log \frac{\exp\left(\boldsymbol{v}_{(i),j}^{c} \cdot \boldsymbol{p}_{(i)}^{c} / \tau\right)}{\sum_{k=1}^{C_i} \exp\left(\boldsymbol{v}_{(i),j}^{c} \cdot \boldsymbol{p}_{(i)}^{k} / \tau\right)}, \tag{4.5}$$

where N represents the number of all skeleton instances, C_i denotes the number of skeleton prototypes generated from the i^{th} instance set $\mathbb{V}_{(i)}$, N_c is the number of instances belonging to the c^{th} prototype $\boldsymbol{p}_{(i)}^{c}$ in $\mathbb{V}_{(i)}$, and τ represents the temperature for contrastive learning. It is worth noting that the naïve prototype contrastive learning (denoted as NPC) using original sequences is a special case of the proposed MPC scheme when $q = 1$ and $x = 0$ (see Eqs. (4.2) and (4.3)). The MPC scheme can be viewed as to perform finer prototype learning with different subsequences, and allow the model to jointly attend to key skeleton patterns from different representation subspaces of the original sequences.

4.2.4 Masked Intra-Sequence Contrastive Learning

The continuity of human motion typically results in very little variation of poses/skeletons within a small temporal interval [4]. Due to this nature, subsequences of the same skeleton sequence usually possess strong inherent correlations. For example, they could locally share similar skeletons and partial sequences with consistent walking patterns. To exploit such intra-sequence relationships and inherent consistency (*e.g.*, pattern invariance) within sequences to learn better skeleton representations, we propose the ***masked intra-sequence contrastive learning (MIC)*** below.

Given two skeleton instances (*i.e.*, subsequence representations), $\boldsymbol{v}_{(i)}$ and $\boldsymbol{v}_{(j)}$, of the same sequence, we first map them into a contrasting space $\mathbb{R}^H$ with a fully-connected (FC) layer $\mathcal{F}_c(\cdot)$ by: $\mathcal{F}_c\left(\boldsymbol{v}_{(i)}\right) = \boldsymbol{z}_{(i)}$ and $\mathcal{F}_c\left(\boldsymbol{v}_{(j)}\right) = \boldsymbol{z}_{(j)}$, where $\boldsymbol{z}_{(i)}, \boldsymbol{z}_{(j)} \in \mathbb{R}^H$. Inspired by [95], we leverage a Siamese architecture to contrast one instance in the original feature space with the other one in the new contrasting space, so as to *symmetrically* learn their inherent similarity. To this end, we exploit a masked intra-sequence contrastive learning (MIC) loss to minimize the negative

cosine similarity between two instances of the same sequence by:

$$\mathcal{L}_{\mathrm{MIC}} = -\alpha \frac{\boldsymbol{z}_{(i)}}{\left\|\boldsymbol{z}_{(i)}\right\|_2} \cdot \frac{\boldsymbol{v}_{(j)}}{\left\|\boldsymbol{v}_{(j)}\right\|_2} - \beta \frac{\boldsymbol{z}_{(j)}}{\left\|\boldsymbol{z}_{(j)}\right\|_2} \cdot \frac{\boldsymbol{v}_{(i)}}{\left\|\boldsymbol{v}_{(i)}\right\|_2}, \tag{4.6}$$

where $\|\cdot\|_2$ denotes ℓ_2-norm, α and β are weights for contrastive learning of representation pairs $(\boldsymbol{z}_{(i)}, \boldsymbol{v}_{(j)})$ and $(\boldsymbol{z}_{(j)}, \boldsymbol{v}_{(i)})$, respectively. Here $\mathcal{L}_{\mathrm{MIC}}$ is defined for two subsequence representations of a skeleton sequence and the total loss is averaged over all sequences. To enable more stable and better contrastive learning, we employ a symmetrized loss with equal weights for two contrastive representation pairs, *i.e.*, $\alpha = \beta = 0.5$, and adopt an alternating stop-gradient operation following [95] when contrasting each pair, as shown in Fig. 4.2 (Note that we only visualize one contrastive pair for conciseness).

4.2.5 Objective Function of SimMC

The proposed SimMC combines both MPC loss (see Eq. (4.5)) and MIC loss (see Eq. (4.6)) to perform unsupervised contrastive learning of skeleton representations with:

$$\mathcal{L} = \lambda \mathcal{L}_{\mathrm{MIC}} + (1 - \lambda)\mathcal{L}_{\mathrm{MPC}}, \tag{4.7}$$

where λ is the weight coefficient to trade off the importance of different contrastive learning. For convenience, here we use $\mathcal{L}_{\mathrm{MIC}}$ to denote the total MIC loss averaging over all training skeleton sequences. To facilitate training and generate more reliable clusters, we optimize our model by alternating clustering and contrastive representation learning. For the person re-ID task, we exploit the encoder $\psi(\cdot)$ learned by our framework to encode each skeleton sequence of the probe set $\Phi_{\mathcal{P}}$ into corresponding representations, $\{\boldsymbol{v}_i^{\mathcal{P}}\}_{i=1}^{N_2}$, which are matched with the representations, $\{\boldsymbol{v}_j^{\mathcal{G}}\}_{j=1}^{N_3}$, of the same identity in the gallery set $\Phi_{\mathcal{G}}$ based on the Euclidean distance.

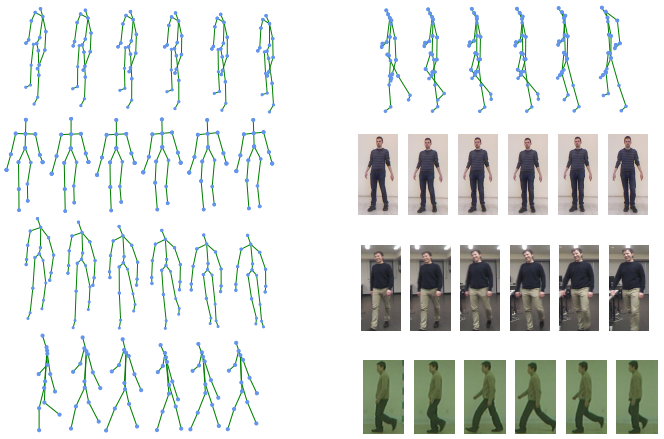


FIGURE 4.3: Examples of 3D skeletons in KGBD (left of 1^{st} row), KS20 (right of 1^{st} row), IAS (2^{nd} row), BIWI (3^{rd} row), and CASIA-B (4^{th} row). We display both skeleton and RGB samples for RGBD datasets (IAS, BIWI).

TABLE 4.1: Overview of datasets.

	KS20	**KGBD**	**IAS**	**BIWI**	**CASIA-B**
# Train IDs	20	164	11	50	124
# Train Skeletons	35,976	188,742	88,986	205,764	706,480
# Gallery IDs	20	164	11	28	62
# Gallery Skeletons	3,252	188,700	A: 6,978 B: 7,764	W: 4,932 S: 3,186	N: 162,080 C: 54,400 B: 53,880
# Probe IDs	20	164	11	28	62
# Probe Skeletons	3,306	94,146	A: 6,978 B: 7,764	W: 4,932 S: 3,186	N: 162,080 C: 54,400 B: 53,880

4.3 Experiments

4.3.1 Experimental Settings

4.3.1.1 Datasets

- **KS20 VisLab Multi-View Kinect Skeleton Dataset (KS20)**[2] [119], which contains multi-view Kinect skeleton (KS) sequences collected from 20 walking subjects using Kinect V2, in the context of long-term person re-identification using biometrics. Multiple walking sequences along five different directions *i.e.*, Left lateral (LL at 0°), Left diagonal (LD at 30°), Frontal

[2]Public Link of KS20 Dataset

(F at 90°), Right diagonal (RD at 130°) and Right lateral (RL at 180°) are collected. Altogether it has 300 skeleton sequences comprising 20 subjects (3 video sequences per person in a particular viewpoint) in the aforementioned directions.

- **Kinect Gait Biometry Dataset (KGBD)**[3] [24], which contains skeleton sequences from 164 subjects extracted using a Kinect sensor. Each subject walked at least once over a semi-circular path and the sensor followed the movement using a spinning dish. All walks were performed indoors with uncontrolled artificial lighting conditions.

- **IAS-Lab RGBD-ID Dataset (IAS)**[4] [120], which is a RGB-D dataset of people targeted to long-term people re-identification from RGB-D cameras. It contains 11 training and 22 testing sequences RGB images, depth images, and 3D skeletons) of 11 different people. For every subject, we recorded three sequences, where the person rotates on himself and performs some walks. The first (IAS training set) and the second sequences (IAS-A testing set) were acquired with people wearing different clothes, while the third one (IAS-B testing set) was collected in a different room, but with the same clothes as in the first sequence. These two different testing sets allow to validate both short-term and long-term re-identification techniques on this dataset.

- **BIWI RGBD-ID Dataset (BIWI)**[5] [1], which contains 50 training and 56 testing sequences (RGB images, depth images, and 3D skeletons) of 50 different people. 28 people out of 50 present in the training set have been recorded also in two testing videos each. These testing sequences have been collected in a different day and in a different location with respect to the training dataset, therefore most subjects are dressed differently. For every person in the testing set, a Still sequence and a Walking sequence have been collected. In the Still video (BIWI-Still), every person is still or slightly moving in place, while in the Walking video (BIWI-Walking), every person performs two walks frontally and two other walks diagonally with respect to the Kinect. Note that our model only uses the 3D skeleton data in IAS and BIWI for training.

[3] Public Link of KGBD Dataset
[4] Public Link of IAS Dataset Link
[5] Public Link of BIWI Dataset Link

- **CASIA-B**[6] [121], which is a large multi-view gait database containing walking sequences (RGB videos) of 124 subjects captured from 11 views (0°, 18°, ..., 180°) and and 3 conditions—pedestrians wearing a bag ("Bags"), wearing a coat ("Clothes"), and without any coat or bag ("Normal"). Three variations, namely view angle, clothing and carrying condition changes, are separately considered. Besides the video files, this dataset also provides human silhouettes extracted from video files. The original CASIA-B dataset does not contain 3D skeleton data, and we follow [25] to exploit pre-trained pose estimation models to extract 3D skeletons from RGB videos of CASIA-B, so as to evaluate the performance of our approach on RGB-estimated skeletons.

4.3.1.2 Probe and Gallery Settings

For the BIWI and IAS datasets, as different testing sets are non-overlapped and contain all pedestrians under different scenes, we evaluate our approach on each testing set by setting it as the probe while the other one is adopted as the gallery. The KGBD dataset contains different skeleton videos (*i.e.*, long skeleton sequences) of each pedestrian with varying numbers of walking rounds. Since no training/testing splits are given, we randomly choose one skeleton video of each person to split skeleton sequences and construct the probe set, and equally divide the remaining videos to build the training set and gallery set. The KS20 dataset collects skeleton data of pedestrians from five different viewpoints, including 0°, 30°, 90°, 130°, and 180°. We randomly take one skeleton sequence from each view as the probe sequence and use one half of the remaining sequences for training and the other half as the gallery. The CASIA-B dataset contains sequences of 124 individuals under 11 different views and 3 conditions—pedestrians wearing a bag ("Bags"), wearing a coat ("Clothes"), and without any coat or bag ("Normal"). We follow the person re-ID protocols in [13] (detailed in Sec. 4.3.1.3) to evaluate the proposed skeleton-based approach on CASIA-B. Experiments with each setup are repeated for multiple times and the average performance is reported in this work.

[6] Public Link of CASIA-B Dataset Link

4.3.1.3 Evaluation Settings of CASIA-B

The 3D skeleton data in existing skeleton-based person re-ID benchmarks (KS20, KGBD, IAS, and BIWI) are collected with Kinect [22]. To evaluate the effectiveness of our approach when 3D skeleton data are directly estimated from RGB videos rather than depth sensors such as Kinect, we use a large-scale RGB video based dataset, *CASIA-B* [121], which contains walking sequences of 124 individuals under 11 different views and 3 conditions—pedestrians wearing a bag ("Bags"), wearing a coat ("Clothes"), and without any coat or bag ("Normal"). We follow the evaluation setup in [13], which is frequently used in the literature: First, we randomly choose half of the individuals for training and use the rest for testing. Then, to evaluate our approach under *single-condition* and *cross-condition* settings, we divide the testing sequences by the three conditions ("Bags", "Clothes", "Normal") to construct gallery or probe sets. Specifically, for the *single-condition* setting, both gallery and probe sets use the testing sequences with the same condition (*i.e.*, gallery and probe sets are the same), and we match each sequence of the probe set with the most similar sequence from the gallery set that *excludes* the original sequence. In the *cross-condition* setting, we adopt the testing sequences under bags ("Bags") or clothes condition ("Clothes") as the probe set, and use the testing sequences under normal condition ("Normal") as the gallery set. Following [25], we exploit pre-trained pose estimation models [122, 123] to extract 3D skeletons from RGB videos of CASIA-B. We first extract eighteen 2D joints from each person in videos using the *OpenPose* model [123]. Then, we follow the same configuration of estimation in [25] and average the positions of "Nose", "Reye", "LEye", "Rear" and "Lear" as the position of "Head" to construct fourteen 2D joints, which are fed into the pose estimation method [122] to estimate corresponding 3D body joints. Thus, J is 14 for CASIA-B, and all joints in each skeleton are normalized by subtracting the neck joint.

4.3.1.4 Implementation Details

Dataset Preprocessing Setups. To avoid ineffective skeleton recording, we discard the first and last 10 skeleton frames of each original skeleton sequence. For KS20, KGBD, BIWI, and IAS datasets, all skeleton sequences are normalized by subtracting the spine joint position from each joint of the same skeleton so that

the skeleton is translation invariant [124]. Then, we spilt all normalized skeleton sequences in the training sets into multiple shorter skeleton sequences (*i.e.*, $\boldsymbol{S}_{1:f}$) with length f by a step of $\frac{f}{2}$, which aims to obtain as many 3D skeleton sequences as possible to train our approach. We split all skeleton sequences in the gallery and probe sets into shorter and non-overlapping sequences with length f. Unless explicitly specified, the skeleton sequence $\boldsymbol{S}_{1:f}$ in our work refers to those split and normalized sequences used in learning, rather than those original skeleton sequences provided by datasets. We follow the data augmentation strategy used in [4, 7, 8] to sample more sequences for different identities in the training set, and train our approach with randomly shuffled and unlabeled skeleton sequences.

Model Parameter Setups. The dimension of each input skeleton is $K = J \times 3$ as we concatenate all 3D coordinates of J body joints in order. The skeleton sequence length f on four skeleton-based datasets (IAS, KS20, BIWI, KGBD) is set to 6 following [4] for a fair comparison with existing methods. We empirically employ $x = 2$ random masks for subsequence sampling, which achieves best performance in average among different settings. As to CASIA-B, it is a large-scale dataset with roughly estimated skeleton data from RGB frames, which is intrinsically different from previous datasets. We adopt a longer sequence length $f = 40$, and empirically set $x = 10$ random masks for subsequence sampling in this dataset. The number of random subsequence sampling is $q = 2$ and the embedding size for skeleton representations is $H = 256$ for all datasets. It should be noted that we set $q = 2$ to generate the instance (*i.e.*, subsequence representation) pair for the MIC scheme, while $q = 1$ is equivalent to $q = 2$ in the MPC scheme as the masking process is random. For DBSCAN clustering in the MPC scheme, we empirically set maximum distance $\epsilon = 0.6$ (KGBD, BIWI-S), $\epsilon = 0.8$ (KS20, IAS-A, IAS-B, BIWI-W), $\epsilon = 0.75$ (CASIA-B), and adopt minimum amount of samples $a_{min} = 4$ for KGBD and $a_{min} = 2$ for other datasets. We follow [29, 125] to construct the commonly used Jaccard distance matrix to perform clustering, and discard all outliers in different clustered instance sets, *i.e.*, discard the union of all outliers, to perform contrastive learning. We set the temperature τ to 0.06 (KGBD), 0.075 (CASIA-B), 0.07 (BIWI), 0.08 (KS20, IAS) for MPC learning. The similarity weights α, β are equally set to 0.5 in the symmetrized MIC loss, while the weight coefficient λ is empirically set to 0.5 (IAS-B, KGBD, CASIA-B), 0.75 (IAS-A), and 0.25 (BIWI). We employ Adam optimizer with learning rate 3.5×10^{-4} for all datasets. The batch size is set to 256 for all datasets. To avoid

over-fitting and achieve better generalization performance, we adopt Early Stopping [126] with a patience of 100 epochs (*i.e.*, stop the training of model after no improvement in 100 continuous epochs). Interested readers can access our source codes at https://github.com/Kali-Hac/SimMC to get more details.

Method Comparison Setups. For all methods compared in our experiments, we select optimal model parameters for training, and use their pre-defined skeleton descriptors or pre-trained skeleton representations for person re-ID. It is worth noting that our re-implementations of some existing models get performance with slight variations, and the results are basically the same or even better than that presented in the original papers under different random model initializations. For direct supervised fine-tuning (DF) of existing models, we attach an MLP network with one hidden layer to the end of original models, and train the MLP network on the frozen pre-trained representations with the supervision of labels. Then the feature representations before the last fully-connected layer are extracted for person re-ID. To perform unsupervised fine-tuning with SimMC, we train SimMC on the *unlabeled* and frozen skeleton representations pre-trained by original models, and exploit the skeleton representations learned by SimMC for person re-ID. Our framework is trained with only unlabeled skeleton data without using any post-processing technique, *e.g.*, re-ranking [127] or multi-query fusion [49]. To perform person re-ID, we exploit the framework to encode each original skeleton sequence without masking (*i.e.*, $x = 0$) of the probe set $\Phi_{\mathcal{P}}$ into corresponding representations, $\{\boldsymbol{v}_i^{\mathcal{P}}\}_{i=1}^{N_2}$, and match it with representations, $\{\boldsymbol{v}_j^{\mathcal{G}}\}_{j=1}^{N_3}$, of the same identity in the gallery set $\Phi_{\mathcal{G}}$ using Euclidean distance. In ablation study, we use the concatenation of raw skeleton sequences (*i.e.*, normalized 3D coordinates of body joints) as the baseline.

4.3.1.5 Evaluation Metrics

We compute the Cumulative Matching Characteristics (CMC) curve and adopt top-1 accuracy (*i.e.*, Rank-1 accuracy), top-5 accuracy, and top-10 accuracy as performance metrics. The top-1 accuracy, top-5 accuracy, and top-10 accuracy are computed as the ratios of probe sequences matching the gallery sequences with correct identities when the candidate gallery sequences are the top 1, top 1 to 5, and top 1 to 10 most similar sequences to the probe sequence. Mean Average

TABLE 4.2: Performance comparison with existing state-of-the-art skeleton-based methods on KS20, KGBD, and IAS-A. "+ DF" denotes direct supervised fine-tuning. **Bold** refers to the best cases among self-supervised/unsupervised methods, while *italics* indicate achieving higher performance when exploiting SimMC ("+ SimMC") to fine-tune corresponding pre-trained representations.

				KS20				KGBD				IAS-A			
Types	**Methods**	**# Params**	**GFLOPs**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**
Hand-crafted	D_{13} [1]	—	—	39.4	71.7	81.7	18.9	17.0	34.4	44.2	1.9	40.0	58.7	67.6	24.5
	D_{16} [2]	—	—	51.7	77.1	86.9	24.0	31.2	50.9	59.8	4.0	42.7	62.9	70.7	25.2
Supervised	PoseGait [25]	8.93M	121.60	49.4	80.9	90.2	23.5	50.6	67.0	72.6	13.9	28.4	55.7	69.2	17.5
	SGELA [4] + DF	9.09M	7.48	49.7	67.0	77.1	22.2	43.7	58.7	65.0	7.1	18.0	32.1	46.2	13.5
	MG-SCR [7]	0.35M	6.60	46.3	75.4	84.0	10.4	44.0	58.7	64.6	6.9	36.4	59.6	69.5	14.1
	SM-SGE [8] + DF	6.25M	23.92	49.8	78.1	85.2	11.7	43.2	58.6	64.6	7.5	38.5	63.2	73.9	15.0
Self-supervised /Unsupervised	AGE [3]	7.15M	37.37	43.2	70.1	80.0	8.9	2.9	5.6	7.5	0.9	31.1	54.8	67.4	13.4
	SGELA [4]	8.47M	7.47	45.0	65.0	75.1	21.2	38.1	53.5	60.0	4.5	16.7	30.2	44.0	13.2
	SM-SGE [8]	5.58M	22.61	45.9	71.9	81.2	9.5	38.2	54.2	60.7	4.4	34.0	60.5	71.6	13.6
	SimMC (Ours)	0.15M	0.99	**66.4**	**80.7**	**87.0**	**22.3**	**54.9**	**66.2**	**70.6**	**11.7**	**44.8**	**65.3**	**72.9**	**18.7**
Unsupervised Fine-tuinig	SGELA + SimMC	8.80M	10.10	*47.3*	*69.7*	*79.3*	20.1	*51.7*	*62.7*	*67.9*	*15.1*	*16.8*	*33.3*	*48.7*	12.0
	MG-SCR + SimMC	0.53M	7.88	*71.1*	*83.6*	*89.1*	*22.7*	*47.4*	*59.3*	*64.9*	*11.0*	*47.2*	*69.0*	*77.3*	*22.4*
	SM-SGE + SimMC	5.89M	25.10	*67.2*	*82.2*	*88.5*	*23.0*	*47.1*	*59.2*	*64.9*	*10.8*	*51.3*	*69.9*	*75.6*	*27.3*

Precision (mAP) [49] is also used to quantitatively evaluate the overall performance of our approach.

4.3.2 Comparison with State-of-the-Art Methods

We compare our framework with existing state-of-the-art self-supervised and unsupervised skeleton-based methods on KS20, KGBD, IAS, and BIWI in Table 4.2 and 4.3. We also include the latest supervised skeleton-based methods and representative hand-crafted methods as a performance reference. The amount of network parameters (million (M)) and computational complexity (giga floating-point operations (GFLOPs)) for different deep learning based methods are reported in Table 4.2.

4.3.2.1 Comparison with Self-supervised and Unsupervised Methods

Our framework shows evident advantages in terms of performance and efficiency over existing state-of-the-art self-supervised and unsupervised methods. As reported in Table 4.2 and 4.3, SimMC significantly outperforms AGE [3] and SM-SGE [8] that manually design pretext tasks based on pre-defined skeleton modeling such as skeleton graphs by a large margin of 7.4-52.0% top-1 accuracy and 2.2-13.4% mAP on all datasets. Compared with the SGELA model [4] using direct

TABLE 4.3: Performance comparison on IAS-B, BIWI-W, BIWI-S. **Bold** refers to the best cases among self-supervised/unsupervised methods, while *italics* indicate achieving higher performance with the fine-tuning of SimMC.

		IAS-B				BIWI-W				BIWI-S			
Types	**Methods**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**
Hand-crafted	D_{13} [1]	43.7	68.6	76.7	23.7	14.2	20.6	23.7	17.2	28.3	53.1	65.9	13.1
	D_{16} [2]	44.5	69.1	80.2	24.5	17.0	25.3	29.6	18.8	32.6	55.7	68.3	16.7
Supervised	PoseGait [25]	28.9	51.6	62.9	20.8	8.8	23.0	31.2	11.1	14.0	40.7	56.7	9.9
	SGELA [4] + DF	23.6	42.9	51.9	14.8	13.9	15.3	16.7	22.9	29.2	65.2	73.8	23.5
	MG-SCR [7]	32.4	56.5	69.4	12.9	10.8	20.3	29.4	11.9	20.1	46.9	64.1	7.6
	SM-SGE [8] + DF	44.3	68.2	77.5	14.9	16.7	31.0	40.2	18.7	34.8	60.6	71.5	12.8
Self-supervised /Unsupervised	AGE [3]	31.1	52.3	64.2	12.8	11.7	21.4	27.3	12.6	25.1	43.1	61.6	8.9
	SGELA [4]	22.2	40.8	50.2	14.0	11.7	14.0	14.7	19.0	25.8	51.8	64.4	**15.1**
	SM-SGE [8]	38.9	64.1	75.8	13.3	13.2	25.8	33.5	15.2	31.3	56.3	69.1	10.1
	SimMC (Ours)	**46.3**	**68.1**	**77.0**	**22.9**	**24.5**	**36.7**	**44.5**	**19.9**	**41.7**	**66.6**	**76.8**	12.3
Unsupervised Fine-tuinig	SGELA + SimMC	21.2	39.1	48.8	14.0	*18.4*	*23.1*	*25.0*	*28.7*	*51.8*	*71.3*	*74.4*	*43.3*
	MG-SCR + SimMC	*52.4*	*72.0*	*78.8*	*29.1*	*25.1*	*37.5*	*46.4*	*20.3*	*28.3*	*51.6*	*64.8*	*10.9*
	SM-SGE + SimMC	*55.3*	*72.6*	*80.3*	*34.1*	*25.9*	*39.2*	*45.2*	*22.4*	*42.6*	*64.8*	*76.2*	*15.4*

inter-sequence contrastive learning, our framework achieves remarkably better performance on five out of six testing sets (KS20, KGBD, IAS-A, IAS-B, BIWI-W) by up to 28.1% top-1 accuracy and 8.9% mAP, which demonstrates that the proposed SimMC combining both prototype (MPC) and intra-sequence contrastive learning (MIC) can capture more discriminative features within skeleton sequences for person re-ID on different datasets. Notably, SimMC also enjoys the smallest model size (only 0.15M) for skeleton representation learning among all approaches shown in Table 4.2, which suggests its higher model efficiency for person re-ID tasks.

By applying the proposed framework to fine-tuning SGELA and SM-SGE models, we can further improve their performance with an average gain of 16.9% and 8.1% top-1 accuracy respectively on all datasets. Such results demonstrate both effectiveness and scalability of proposed masked contrastive learning, which is compatible with existing models and can fully exploit their pre-trained features to achieve higher-quality skeleton representations for person re-ID.

4.3.2.2 Comparison with Hand-crafted and Supervised Methods

In contrast to hand-crafted methods (D_{13} and D_{16}) that rely on geometric joint distances and anthropometric descriptors, our approach obtains similar performance on IAS testing sets, while it achieves a distinct improvement of 7.5-37.9% top-1 accuracy on BIWI, KS20, and KGBD datasets that contain more views and individuals. Despite utilizing *unlabeled* skeleton data as the sole input, the proposed SimMC still performs better than the latest supervised models PoseGait and

MG-SCR in most cases. Interestingly, applying SimMC to SM-SGE achieves significantly higher performance gains than direct supervised fine-tuing (DF) in terms of top-1 accuracy (3.9-17.4%), top-5 accuracy (0.6-6.7%), top-10 accuracy (0.3-5.0%), and mAP (3.3-19.2%) on all datasets. With highly efficient performance and strong scalability, the proposed unsupervised SimMC can be a more general framework for skeleton-based person re-ID and related tasks.

4.4 Further Analysis

4.4.1 Ablation Study

We conduct ablation study to demonstrate the contribution of each component in our framework, including masked prototype contrastive learning (MPC) scheme, and masked intra-sequence contrastive learning (MIC), and compare them with Naïve prototype contrastive learning (NPC) using only original sequences. We adopt 3D coordinates of raw skeleton sequences as the baseline representation for person re-ID. As reported in Table 4.4, the model exploiting NPC significantly outperforms the baseline by 9.8-47.8% top-1 accuracy and 2.0-11.0% mAP. Considering that NPC is a special case of the proposed MPC scheme (see Sec. 4.2.3), such results verify the effectiveness of the skeleton prototype contrastive learning in MPC, which can capture highly discriminative features within unlabeled skeleton sequences for the task of person re-ID. Employing the standard MPC scheme with randomly sampled subsequences consistently improves the model performance by up to 3.9% top-1 accuracy and 1.2% mAP on all datasets, which demonstrates that MPC is able to mine more representative key features from skeleton subsequences to perform person re-ID. Finally, incorporating MIC into MPC further improves model performance with 0.8-2.5% top-1 accuracy and 0.2%-1.2% mAP gains on different datasets. This justifies our claim that capturing inherent intra-sequence similarity and pattern consistency within sequences could facilitate learning better representations of skeleton sequences for person re-ID.

TABLE 4.4: Ablation study of SimMC framework with different configurations: Naïve prototype contrastive learning (NPC) using only original sequences, masked prototype contrastive learning (MPC) scheme and corresponding masked intra-sequence contrastive learning (MIC).

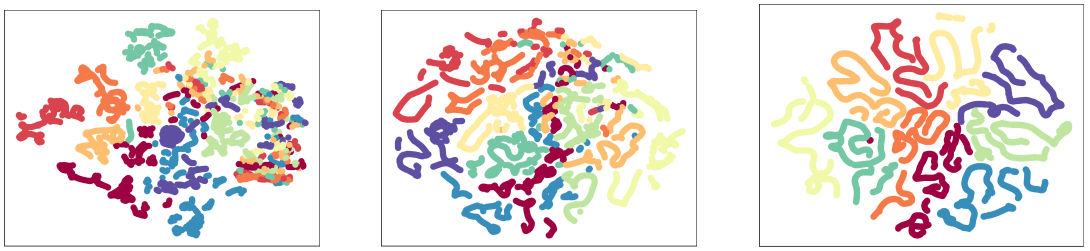
	IAS-A		IAS-B		BIWI-S		BIWI-W		KS20		KGBD	
Configurations	**top-1**	**mAP**	**top-1**	**mAP**	**top-1**	**mAP**	**top-1**	**mAP**	**top-1**	**mAP**	**top-1**	**mAP**
Baseline	29.4	13.8	30.2	13.3	24.8	9.3	10.9	14.1	17.0	9.5	34.5	6.4
NPC	39.2	17.8	40.7	21.5	38.1	11.3	21.2	18.3	64.8	20.5	53.0	11.0
MPC	43.1	18.5	43.8	22.3	40.1	11.7	23.7	19.5	65.6	21.1	53.6	11.0
MPC + MIC	44.8	18.7	46.3	22.9	41.7	12.3	24.5	19.9	66.4	22.3	54.9	11.7

TABLE 4.5: Comparison with appearance-based and skeleton-based methods on CASIA-B. "Bg-Nm" represents the probe set with "Bags (Bg)" condition and gallery set with "Normal (Nm)" condition. "—" indicates no published result.

Probe-Gallery	**Nm-Nm**				**Bg-Bg**				**Cl-Cl**				**Cl-Nm**				**Bg-Nm**			
Methods	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**
LMNN [9]	3.9	22.7	36.1	—	18.3	38.6	49.2	—	17.4	35.7	45.8	—	11.6	12.6	17.8	—	23.1	37.1	44.4	—
ITML [10]	7.5	22.2	34.2	—	19.5	26.0	33.7	—	20.1	34.4	43.3	—	10.3	24.5	36.1	—	21.8	30.4	36.3	—
ELF [11]	12.3	35.6	50.3	—	5.8	25.5	37.6	—	19.9	43.9	56.7	—	5.6	16.0	26.3	—	17.1	30.0	37.9	—
SDALF [12]	4.9	27.0	41.6	—	10.2	33.5	47.2	—	16.7	42.0	56.7	—	11.6	19.4	27.6	—	22.9	30.1	36.1	—
Score-based MLR [13]	13.6	48.7	63.7	—	13.6	48.7	63.7	—	13.5	48.6	63.9	—	9.7	27.8	45.1	—	14.7	32.6	50.2	—
Feature-based MLR [13]	16.3	43.4	60.8	—	18.9	44.8	59.4	—	25.4	53.3	68.9	—	20.3	42.6	**56.9**	—	31.8	53.6	64.1	—
AGE [3]	20.8	29.3	34.2	3.5	37.1	56.2	67.0	9.8	35.5	54.3	65.3	9.6	14.6	33.0	42.7	3.0	32.4	51.2	60.1	3.9
SM-SGE [8]	50.2	73.5	81.9	6.6	26.6	49.0	59.4	9.3	27.2	51.4	63.2	9.7	10.6	26.3	35.9	3.0	16.6	36.8	47.5	3.5
SGELA [4]	71.8	87.5	91.4	9.8	48.1	69.5	77.7	**16.5**	51.2	73.8	81.5	7.1	15.9	30.8	40.6	4.7	36.4	57.1	64.6	6.7
SimMC (Ours)	**84.8**	**92.3**	**93.7**	**10.8**	**69.1**	**86.6**	**91.3**	**16.5**	**68.0**	**88.1**	**93.0**	**15.7**	**25.6**	**43.8**	54.0	**5.4**	**42.0**	**59.8**	**68.9**	**7.1**

4.4.2 Evaluation on RGB-Estimated Skeletons

To verify the effectiveness of SimMC when applied to RGB-based scenes with model-estimated 3D skeletons, we utilize pre-trained pose estimation models to extract skeleton data from RGB videos of CASIA-B, and compare the performance of SimMC with representative appearance-based and skeleton-based methods. As reported in Table 4.5, the proposed SimMC remarkably outperforms state-of-the-art skeleton-based models SM-SGE and SGELA by a distinct margin of 5.6% to 42.5% top-1 accuracy and 0.4% to 8.6% mAP in different conditions, which suggests the stronger ability of our framework on capturing discriminative features from estimated skeleton data for person re-ID. Compared with appearance-based ELF and MLR models that utilize visual features (*e.g.*, colors, textures, and silhouettes) with extra label information, the skeleton-based SimMC also achieves superior performance in all conditions of CASIA-B, which demonstrates its great applicable value and potential for person re-ID under large-scale RGB-based scenarios and more general settings.



(A) AGE (B) SM-SGE (C) SimMC (Ours)

FIGURE 4.4: t-SNE visualization of representations learned by AGE (a), SM-SGE (b), and SimMC (c) for first ten classes in BIWI. Different colors denote skeleton representations of different classes.

TABLE 4.6: Performance of SimMC framework on different datasets when setting different weight coefficients ($\lambda = 0.00, 0.25, 0.50, 0.75, 1.00$)

	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
λ	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
0.00	65.6	21.1	53.6	11.0	43.1	18.5	43.8	22.3	23.7	19.5	40.1	11.7
0.25	67.0	21.4	55.7	11.6	43.3	18.0	44.6	21.7	24.5	19.9	41.7	12.3
0.50	66.4	22.3	54.9	11.7	43.3	18.4	46.3	22.9	25.1	20.4	39.7	11.5
0.75	66.2	22.1	54.7	11.7	44.8	18.7	45.7	22.2	25.5	19.6	41.0	12.2
1.00	51.4	11.3	46.8	4.9	37.5	17.1	39.9	16.2	16.6	15.0	34.8	9.1

4.4.3 Feature Visualization

As shown in Fig. 4.4, we conduct a t-SNE visualization [128] of skeleton representations learned by AGE [3], SM-SGE [8], and the proposed SimMC. The skeleton representations learned by our framework are clustered with higher inter-class separation than AGE and SM-SGE, which suggests that SimMC may learn richer class-related semantics and lower-entropy skeleton representations.

4.4.4 Analysis of Hyperparameters

We intuitively show effects of different hyperparameters on SimMC in Fig. 4.5. The results indicate that the use of random masks ($x > 0$) is the key to the proposed masked contrastive learning, regardless of adopting unified or non-unified mask numbers, while an appropriate fusion ($\lambda > 0$) of MIC and MPC facilitates better skeleton representation learning for person re-ID. Our framework with the optimal parameter setting is not sensitive to changes of some parameters such as temperatures τ. We systematically analyze the effects of hyper-parameters as follows.

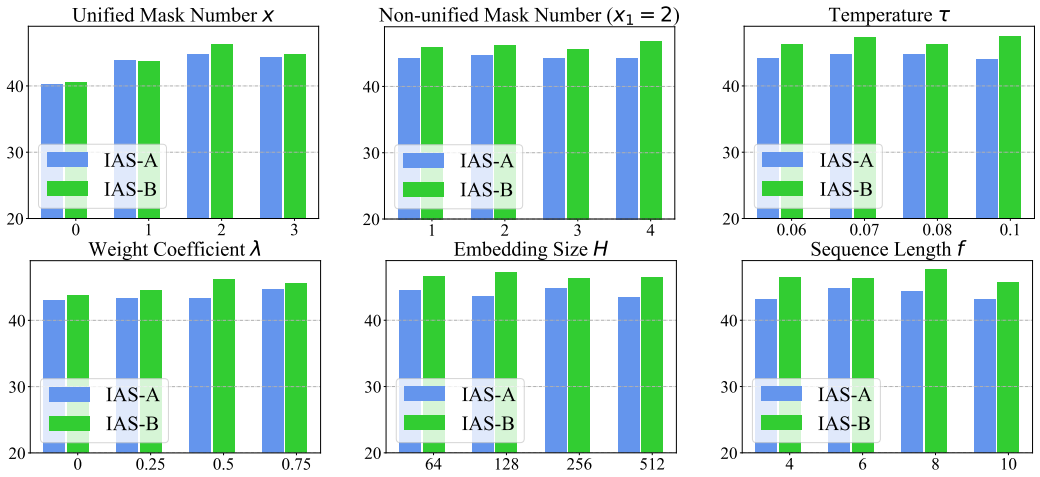


FIGURE 4.5: Top-1 accuracy on IAS-A/B showing effects of hyper-parameters. "Non-unified Mask Number ($x_1 = 2$)" denotes using different mask numbers including $x = 2$ for subsequence sampling.

TABLE 4.7: Performance of SimMC framework on different datasets when setting different numbers of masks ($x = 0, 1, 2, 3, 4$)

	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
x	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
0	65.8	20.8	54.3	10.9	40.2	17.3	40.6	21.3	23.4	18.7	39.8	11.6
1	66.2	20.7	54.9	11.0	43.8	18.2	43.7	21.5	23.6	20.1	41.5	12.2
2	66.4	22.3	54.9	11.7	44.8	18.7	46.3	22.9	24.5	19.9	41.7	12.3
3	66.6	21.6	54.9	11.3	44.3	17.8	44.8	22.1	24.9	18.8	41.0	12.1
4	63.9	22.0	54.7	12.0	42.4	17.1	41.3	17.3	20.7	16.8	41.2	11.7

TABLE 4.8: Performance of SimMC framework on different datasets when setting different minimum amounts of samples for the DBSCAN algorithm ($a_{min} = 1, 2, 3, 4$).

	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
a_{min}	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
1	66.3	22.4	56.4	11.4	43.0	18.7	46.6	22.5	24.7	19.8	41.6	12.0
2	66.4	22.3	55.8	11.7	44.8	18.7	46.3	22.9	24.5	19.9	41.7	12.3
3	67.4	22.3	54.7	11.9	44.4	18.3	47.9	22.9	24.6	19.9	42.2	11.6
4	65.6	21.3	54.9	11.7	44.8	19.9	47.0	23.7	25.4	20.4	42.6	12.3

TABLE 4.9: Performance of SimMC framework on different datasets when setting different maximum distances ($\epsilon = 0.4, 0.6, 0.8, 1.0$).

	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
ϵ	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
0.4	65.8	22.9	52.1	9.2	44.5	20.2	47.4	25.6	17.5	16.3	39.8	11.2
0.6	67.0	22.5	54.9	11.7	43.3	19.6	46.2	23.7	23.2	18.5	41.7	12.3
0.8	66.4	22.3	48.5	5.3	44.8	18.7	46.3	22.9	24.5	19.9	40.5	12.1
1.0	66.8	21.6	44.6	5.0	42.6	18.5	43.6	22.3	20.2	17.3	24.8	7.1

4.4.4.1 Coefficient Settings for Contrastive Learning

We evaluate the performance of our framework with different weight coefficients ($\lambda = 0.00, 0.25, 0.50, 0.75, 1.00$) for combining MPC and MIC. As shown in Table 4.6, the proposed SimMC combining MPC and MIC achieves higher performance

TABLE 4.10: Performance of SimMC framework on different datasets when setting different numbers of hidden layers for the MLP encoder.

	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
Hidden Layers	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
1	66.4	22.3	54.9	11.7	44.8	18.7	46.3	22.9	24.5	19.9	41.7	12.3
2	64.8	21.4	52.2	11.0	41.6	18.8	46.2	23.0	23.2	18.9	40.4	11.6
3	64.7	21.7	48.4	9.9	40.4	18.1	42.8	22.3	22.8	19.5	38.3	12.0
4	62.7	22.0	46.2	9.3	39.8	18.9	40.9	22.7	19.5	15.6	34.2	12.6

TABLE 4.11: Performance of SimMC framework on different datasets when adopting different types of predictor head (fully-connected layer (FC) or multi-layer perceptron (MLP)).

	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
Predictor Head	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
FC	66.4	22.3	54.9	11.7	44.8	18.7	46.3	22.9	24.5	19.9	41.7	12.3
MLP	66.6	21.9	55.8	11.2	44.7	19.1	48.1	24.9	24.6	20.7	42.7	11.9

than solely employing MPC ($\lambda = 0.00$) or MIC ($\lambda = 1.00$). Since skeleton sequences of different datasets are typically captured or estimated in different conditions (*e.g.*, capturing frequency), the properties of sequences and their inherent relationships might be very different between two datasets, leading to a specific requirement of λ for masked intra-sequence contrastive learning. It can also be inferred that the MPC ($\lambda = 0.00$) contributes more than MIC ($\lambda = 1.00$) to the proposed SimMC, as using only MIC ($\lambda = 1.00$) cannot achieve satisfactory performance on all datasets, while incorporating them can maximize the performance gains. We employ $\lambda = 0.5$ for KS20, KGBD and IAS-B, $\lambda = 0.75$ for IAS-A, $\lambda = 0.25$ for BIWI-W and BIWI-S, which achieves the best performance among different settings.

4.4.4.2 Mask Settings for Subsequence Sampling

We evaluate the performance of SimMC when setting different numbers of masks ($x = 0, 1, 2, 3, 4$) for sampling random subsequences. As shown in Table 4.7, applying masked contrastive learning ($x > 0$) can encourage the model to learn better skeleton representations and achieve higher performance in average than directly performing contrastive learning with original sequences ($x = 0$) and MIC. However, too many masks (*e.g.*, $x > 3$) could hurt the overall model performance on all datasets. Considering that the sequence length of the original sequence is 6, when $x > 3$, more than half of the skeleton sequence and corresponding pattern information are masked and discarded in the training, which could result in a large loss of valuable skeleton features in clustering/contrastive learning and negatively

influence the quality of learned skeleton representations for person re-ID. In our framework, we empirically set a relatively small value $x = 2$ for all datasets.

4.4.4.3 Different Settings of DBSCAN

We provide the performance results of our framework when setting different minimum sample amounts ($a_{min} = 1, 2, 3, 4$) and maximum distances ($\epsilon = 0.4, 0.6, 0.8, 1.0$) in Table 4.8 and Table 4.9, respectively. Note that we adjust a specific parameter while keeping other parameters unchanged. Our framework is robust to the change of a_{min}, while it is more sensitive to the parameter ϵ. Too high ϵ (*e.g.*, $\epsilon = 1.0$) is shown to reduce the overall performance on five of six testing sets (KGBD, IAS-A, IAS-B, BIWI-W, BIWI-S), while adopting relatively small ϵ (*e.g.*, $\epsilon = 0.6$), *i.e.*, reducing the connectedness of skeleton instances in clusters and improving the amount of prototypes, can facilitate person re-ID performance in most cases. Such results suggest that performing masked contrastive learning with more diverse skeleton prototypes could encourage mining richer discriminative features from unlabeled skeletons of different datasets.

4.4.4.4 Different Hidden Layers and Predictor Heads

We show the performance of our framework when exploiting the MLP encoder with different numbers of hidden layers $(1, 2, 3, 4)$ in Table 4.10. It is observed that the addition of hidden layers does not bring extra performance gains to our framework, and it could even lead to an evident performance degradation. Different from [95] that employs two hidden layers with up to 2048-d embedding size to learn large-scale image data, our framework can learn better skeleton representations with simpler network structure on relatively limited training data (Note that ImageNet data used in [95] are much larger than skeleton data and require deeper neural networks to sufficiently learn patterns). Based on this observation, we exploit the MLP encoder with one hidden layer for masked contrastive learning. Likewise, in the Siamese architecture of MIC, we construct a simpler predictor head with fully-connected (FC) layer instead of MLP (see Sec. 4.5.2). As shown in Table 4.11, the predictor head using FC or MLP can achieve highly similar performance on different datasets, while FC can enjoy lower complexity than MLP to construct more efficient framework.

TABLE 4.12: Performance of SimMC framework on different datasets when setting different temperatures for the MPC scheme ($\tau = 0.06, 0.07, 0.08, 0.1, 0.5$).

	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
τ	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
0.06	67.4	21.8	54.9	11.7	44.1	18.6	46.3	21.8	25.1	20.9	41.6	11.9
0.07	66.6	22.2	55.1	12.1	44.8	18.5	47.3	22.2	24.5	19.9	41.7	12.3
0.08	66.4	22.3	55.9	12.4	44.8	18.7	46.3	22.9	24.6	19.8	41.8	10.7
0.1	65.8	21.8	55.3	11.8	44.0	18.3	47.5	23.8	24.1	19.5	41.4	11.2
0.5	67.6	21.4	55.4	12.1	43.9	18.7	48.6	25.1	24.7	19.6	42.0	11.5

TABLE 4.13: Performance of SimMC framework on different datasets when setting different sequence length ($f = 2, 4, 6, 8$).

	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
f	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
4	62.6	19.0	53.9	10.9	43.2	19.2	46.5	20.5	21.2	16.9	38.4	11.1
6	66.4	22.3	54.9	11.7	44.8	18.7	46.3	22.9	24.5	19.9	41.7	12.3
8	69.9	28.0	55.9	12.9	44.4	24.1	47.7	26.7	24.4	27.6	42.6	14.9
10	69.5	24.8	55.5	12.3	43.2	19.5	44.2	30.0	28.9	23.1	37.7	16.5

TABLE 4.14: Performance of SimMC framework on different datasets when setting different embedding sizes for skeleton representations ($H = 64, 128, 256, 512$).

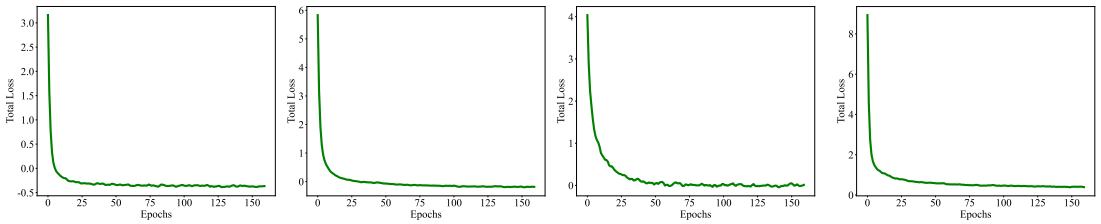
	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
H	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
64	65.4	20.7	51.8	11.2	44.6	18.5	46.7	22.1	15.9	15.9	42.1	12.6
128	67.0	19.8	53.6	11.4	43.7	18.8	47.3	23.2	25.0	20.4	43.0	11.3
256	66.4	22.3	54.9	11.7	44.8	18.7	46.3	22.9	24.5	19.9	41.7	12.3
512	66.2	22.1	56.0	12.1	43.5	18.8	46.5	23.1	25.0	20.6	41.6	11.8

4.4.4.5 Other Hyper-Parameters

Temperature Setting for Contrastive Learning. As shown in Table 4.12, our framework is not sensitive to the changes of temperatures and can achieve comparable performance in the range of 0.06 to 0.5. In practice, we empirically select an appropriate temperature for each dataset to achieve more balanced and better model performance.

Different Sequence Lengths. We provide the performance results of our framework when using different sequence length ($f = 4, 6, 8, 10$) in Table 4.13. In our work, we evaluate all compared methods under the same sequence length ($f = 6$).

Different Embedding Sizes. We evaluate the performance of SimMC when using different embedding sizes ($H = 64, 128, 256, 512$) in Table 4.14. Too small embedding size ($H = 64$) is shown to obtain lower performance in most cases, while larger sizes ($H > 256$) cannot further improve the performance. This suggests that too high-dimensional feature space might be hard to optimize and contain



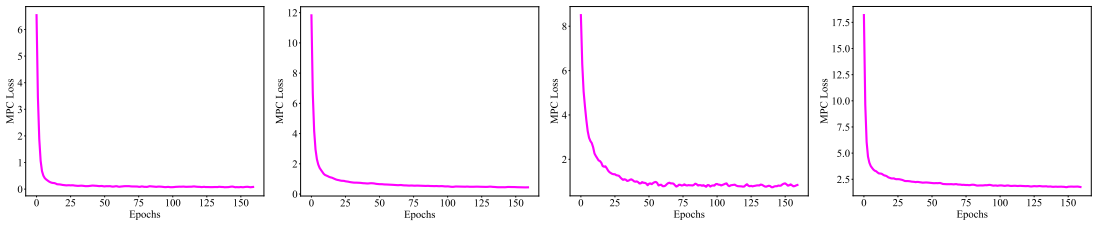
(A) KS20 (B) KGBD (C) IAS (D) BIWI

FIGURE 4.6: Total training losses ($\mathcal{L} = \lambda\mathcal{L}_{\text{MIC}} + (1-\lambda)\mathcal{L}_{\text{MPC}}$) of masked contrastive learning on different datasets.

more redundant information, which cannot benefit learning a compact and effective representation.

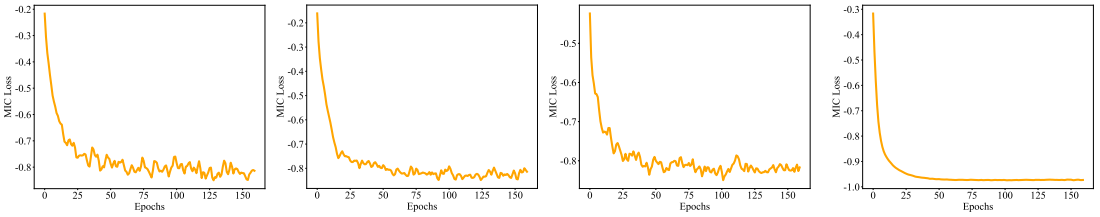
4.4.5 Analysis of Training Process

We visualize the total training loss, MPC training loss, and MIC training losses in Fig. 4.6, Fig. 4.7, and Fig. 4.8 respectively, which show that the training of our framework can converge very fast in the first 50 optimization epochs. To provide a further analysis of the learned skeleton representations, we follow the method in [129] to estimate the mutual information between the skeleton instance features and the ground-truth class labels, as shown in Fig. 4.9. The results show that the training of our framework rapidly and significantly improves the mutual information *w.r.t.* skeleton representations, *i.e.*, similarity between the pseudo classes generated by SimMC and ground-truth class labels, which demonstrates that the proposed masked contrastive learning can encourage the model to capture class-related semantics (*e.g.*, inter-class differences) to learn more discriminative skeleton representations. Besides, we evaluate the uniformity of the representation distributions on the output unit hypersphere, which is one of the key properties related to contrastive learning and can indicate the quality of learned features. We follow [86] to compute the uniform loss in the training process. As shown in Fig. 4.10, the training of SimMC evidently improves the feature uniformity (*i.e.*, lower uniform loss) on different datasets, which suggests that the proposed masked contrastive learning could effectively mine more useful features and enhance the expressiveness of skeleton representations to [130].



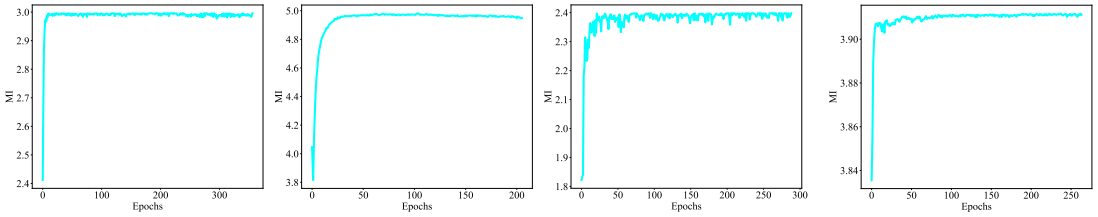
(A) KS20 (B) KGBD (C) IAS (D) BIWI

FIGURE 4.7: Training losses ($\mathcal{L}_{\text{MPC}}$) of masked prototype contrastive (MPC) learning on different datasets.



(A) KS20 (B) KGBD (C) IAS (D) BIWI

FIGURE 4.8: Training losses ($\mathcal{L}_{\text{MIC}}$) of masked intra-sequence contrastive (MIC) learning on different datasets.

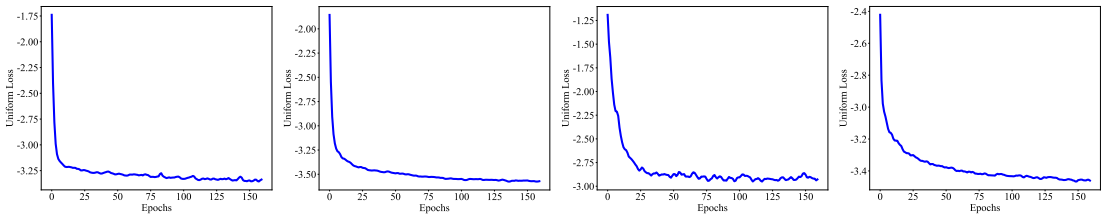


(A) KS20 (B) KGBD (C) IAS (D) BIWI

FIGURE 4.9: Mutual information (MI) between the clusters/pseudo classes generated by SimMC and ground-truth class labels on the training sets of different datasets.

4.4.6 Analysis of Confusion Matrix

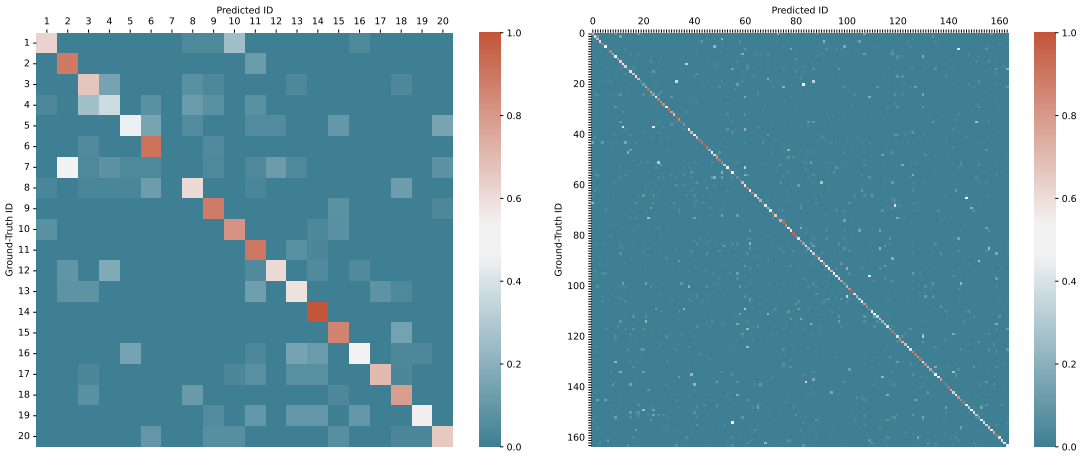
In Fig. 4.11, we show the confusion matrices of our framework when performing person re-ID with the Rank-1 matching (*i.e.*, predicting the identity of each probe sequence using the top-1 gallery sequence that has the smallest Euclidean distance) on all testing sets (probe sets). Note that abscissa and ordinate in Fig. 4.11 denote the predicted and ground-truth identities, respectively. The position in the a^{th} row and b^{th} column indicates that the testing samples belonging to the a^{th} identity is predicted as the b^{th} identity, while the corresponding value is the proportion of such samples to the same-class samples in the testing set. As presented in Fig. 4.11 (a)-(f), the numbers of classes with high accuracy (*i.e.*, number of red grids on the



(A) KS20 (B) KGBD (C) IAS (D) BIWI

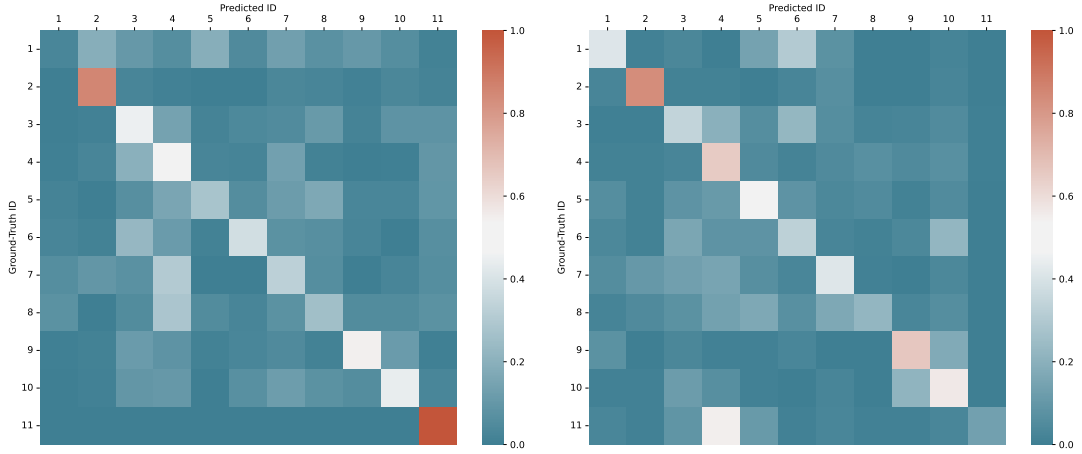
FIGURE 4.10: Uniform loss curves on the training sets of different datasets. Lower uniform loss indicates higher uniformity of feature distributions on the output unit hypersphere, which could suggest the higher quality of the learned representations (see Sec. 4.4.5).

diagonal line) in KS20, KGBD, IAS-A, IAS-B, and BIWI-Still are larger than that in BIWI-Walking. Intuitively, the larger numbers of white and red grids diffused around the diagonal lines, which represent the higher proportions of false matches, on the matrix of BIWI-Walking (see Fig. 4.11 (f)) imply that our model tends to confuse skeleton sequences of more different identities on this dataset. These results are consistent with the performance results shown in Sec. 4.3.2.



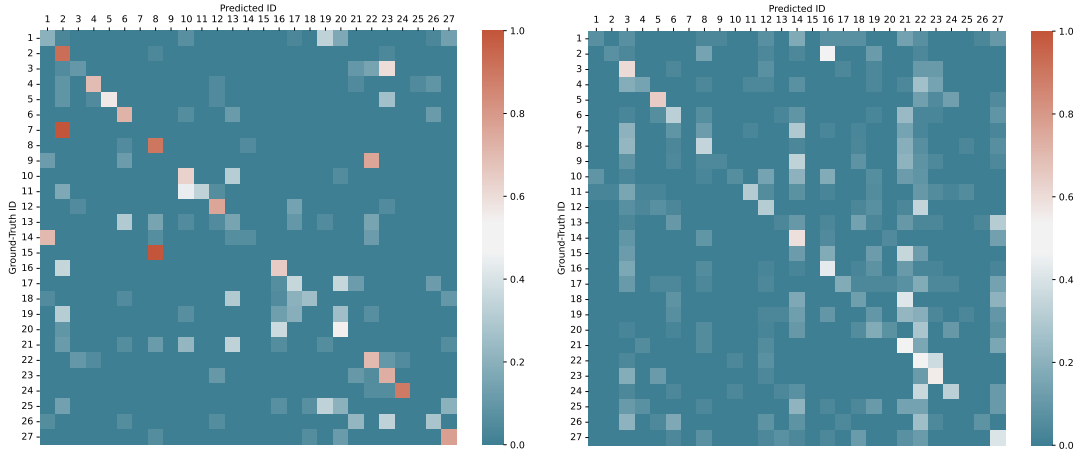
(A) Confusion Matrix on KS20

(B) Confusion Matrix on KGBD



(C) Confusion Matrix on IAS-A

(D) Confusion Matrix on IAS-B



(E) Confusion Matrix on BIWI-Still

(F) Confusion Matrix on BIWI-Walking

FIGURE 4.11: Visualization of confusion matrices on KS20 (a), KGBD (b), IAS-A (c), IAS-B (d), BIWI-Still (e), and BIWI-Walking (f) when using the top-1 matching.

4.5 Theoretical Hypotheses and Analyses

The proposed masked prototype contrastive learning (MPC) and masked intra-sequence contrastive learning (MIC) can be formulated as Expectation-Maximization (EM) solutions. In this section, we provide a theoretical EM modeling for each component of the proposed framework to prove its effectiveness and convergence.

4.5.1 MPC Modeled as Expectation-Maximization Algorithm

Preliminaries. For clarity and convenience, we adopt a more general notation here, which is different from those used in the previous sections. Suppose that a training set $X = \{\boldsymbol{x}_i\}_{i=1}^N$ contains N skeleton sequences, where $\boldsymbol{x}_i \in \mathbb{R}^{f \times K}$, the objective of unsupervised skeleton representation learning is to learn an embedding/encoder function f_θ (realized via θ-parameterized neural networks) that maps X to $V = \{\boldsymbol{v}_i\}_{i=1}^N$, where $\boldsymbol{v}_i \in \mathbb{R}^H$, by $\boldsymbol{v}_i = f_\theta(\boldsymbol{x}_i)$ without using any label, such that $\boldsymbol{v}_i$ can effectively represent features of $\boldsymbol{x}_i$ to perform person re-identification.

Formally, the goal is to find the network parameter θ that maximizes the log-likelihood function of the observed N skeleton sequences as follows:

$$\begin{aligned}\theta^* &= \arg\max_\theta L(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N; \theta) \\ &= \arg\max_\theta \prod_{i=1}^N p(\boldsymbol{x}_i; \theta) \iff \arg\max_\theta \sum_{i=1}^N \log p(\boldsymbol{x}_i; \theta),\end{aligned} \tag{1}$$

where $L(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N; \theta)$ denotes the likelihood function of the observed skeleton sequences $\{\boldsymbol{x}_i\}_{i=1}^N$ *w.r.t* θ, and each skeleton sequence $\boldsymbol{x}_i$ is hypothetically related to a certain skeleton prototype $\boldsymbol{c}_j \in \mathbb{R}^H$ and $\boldsymbol{c}_j \in \{\boldsymbol{c}_j\}_{j=1}^K$. Under this assumption, we can re-formulate the objective in Eq. (1) as:

$$\begin{aligned}\theta^* =& \arg\max_\theta \sum_{i=1}^N \log p(\boldsymbol{x}_i; \theta) \\ =& \arg\max_\theta \sum_{i=1}^N \log \sum_{j=1}^K p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta),\end{aligned} \tag{2}$$

Directly optimizing this function is intractable, thus we consider a lower-bound by using a surrogate function as:

$$\begin{aligned}\sum_{i=1}^{N}\log\sum_{j=1}^{K}p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right)&=\sum_{i=1}^{N}\log\sum_{j=1}^{K}Q\left(\boldsymbol{c}_j\right)\frac{p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right)}{Q\left(\boldsymbol{c}_j\right)}\\&\geq\sum_{i=1}^{N}\sum_{j=1}^{K}Q\left(\boldsymbol{c}_j\right)\log\frac{p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right)}{Q\left(\boldsymbol{c}_j\right)},\end{aligned}\tag{3}$$

where $Q\left(\boldsymbol{c}_j\right)$ represents some distribution over $\{\boldsymbol{c}_j\}_{j=1}^{K}$ and $\sum_{j=1}^{K}Q\left(\boldsymbol{c}_j\right)=1$. We apply Jensen's inequality to derive the last step, where equality can be achieved under the condition that $\frac{p(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta)}{Q(\boldsymbol{c}_j)}$ is a constant. To realize this equality, we have:

$$Q\left(\boldsymbol{c}_j\right)=\frac{p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right)}{\sum_{j=1}^{K}p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right)}=\frac{p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right)}{p\left(\boldsymbol{x}_i;\theta\right)}=p\left(\boldsymbol{c}_j;\boldsymbol{x}_i,\theta\right),\tag{4}$$

where $Q\left(\boldsymbol{c}_j\right)$ is a posterior probability related to $\boldsymbol{c}_j,\boldsymbol{x}_i$, and θ. When θ is fixed at the Expectation step, the distribution of representations $(\boldsymbol{x}_i)$ and corresponding prototypes $(\boldsymbol{c}_j)$ can be estimated as a result of clustering, thus we can get the constant value of $Q\left(\boldsymbol{c}_j\right)=p\left(\boldsymbol{c}_j;\boldsymbol{x}_i,\theta\right)$ based on the result. We can re-write Eq. (3) as:

$$\sum_{i=1}^{N}\sum_{j=1}^{K}\left(Q\left(\boldsymbol{c}_j\right)\log p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right)-Q\left(\boldsymbol{c}_j\right)\log Q\left(\boldsymbol{c}_j\right)\right),\tag{5}$$

where the constant $-\sum_{i=1}^{N}\sum_{j=1}^{K}Q\left(\boldsymbol{c}_j\right)\log Q\left(\boldsymbol{c}_j\right)$ can be ignored and we need to maximize:

$$\sum_{i=1}^{N}\sum_{j=1}^{K}Q\left(\boldsymbol{c}_j\right)\log p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right)\tag{6}$$

For the **Expectation (E)-step**, we aim to estimate $p\left(\boldsymbol{c}_j;\boldsymbol{x}_i,\theta\right)$ (see Eq. (4)). In our framework, we run the DBSCAN algorithm on the features $\boldsymbol{v}_i=f_\theta\left(\boldsymbol{x}_i\right)$ given by the encoder to obtain K clusters $\{\boldsymbol{C}_j\}_{j=1}^{K}$. We generate corresponding skeleton prototype $\boldsymbol{c}_j$, which is the centroid of the j^{th} cluster $\boldsymbol{C}_j$. Then, we compute $p\left(\boldsymbol{c}_j;\boldsymbol{x}_i,\theta\right)=\mathbb{1}\left(\boldsymbol{x}_i\in\boldsymbol{C}_j\right)$, where $\mathbb{1}\left(\boldsymbol{x}_i\in\boldsymbol{C}_j\right)=1$ if $\boldsymbol{x}_i$ belongs to the j^{th} cluster $\boldsymbol{C}_j$ (*i.e.*, corresponding to skeleton prototype $\boldsymbol{c}_j$); otherwise $\mathbb{1}\left(\boldsymbol{x}_i\in\boldsymbol{C}_j\right)=0$.

Assumption 4.1. ***Prototype-Cluster Consistency.*** The global distribution of prototypes is consistent with the distribution of cluster centroids, *i.e.*, each cluster explicitly corresponds to the group of instances that belong to the same prototype. In the E-step, we adopt this commonly-used assumption [5, 88] to generate skeleton prototypes and derive $p\left(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta\right) = \mathbb{1}\left(\boldsymbol{x}_i \in \boldsymbol{C}_j\right)$.

In the **Maximization (M)-step**, we combine Eq. (4) to maximize the lower-bound in Eq. (6) after the E-step:

$$\begin{aligned} &\sum_{i=1}^{N}\sum_{j=1}^{K} Q\left(\boldsymbol{c}_j\right) \log p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right) \\ =&\sum_{i=1}^{N}\sum_{j=1}^{K} p\left(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta\right) \log p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right) \\ =&\sum_{i=1}^{N}\sum_{j=1}^{K} \mathbb{1}\left(\boldsymbol{x}_i \in \boldsymbol{C}_j\right) \log p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right) \end{aligned} \tag{7}$$

Each cluster centroid $\boldsymbol{c}_j$ is assumed to have a uniform prior probability $p\left(\boldsymbol{c}_j; \theta\right) = \frac{1}{K}$ since we are not provided any samples. We have:

$$p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right) = p\left(\boldsymbol{x}_i; \boldsymbol{c}_j, \theta\right) p\left(\boldsymbol{c}_j; \theta\right) = \frac{1}{K} \cdot p\left(\boldsymbol{x}_i; \boldsymbol{c}_j, \theta\right), \tag{8}$$

where the distribution of samples around each prototype is assumed as an isotropic Gaussian, leading to:

$$p\left(\boldsymbol{x}_i; \boldsymbol{c}_j, \theta\right) = \frac{\exp\left(\frac{-\left(\boldsymbol{v}_i - \boldsymbol{c}_p\right)^2}{2\sigma_p^2}\right)}{\sum_{j=1}^{K} \exp\left(\frac{-\left(\boldsymbol{v}_i - \boldsymbol{c}_j\right)^2}{2\sigma_j^2}\right)}, \tag{9}$$

where $\boldsymbol{v}_i = f_\theta\left(\boldsymbol{x}_i\right)$ and $\boldsymbol{c}_p$ is the prototype for the cluster $\boldsymbol{C}_p$ containing $\boldsymbol{x}_i$, *i.e.*, $\boldsymbol{x}_i \in \boldsymbol{C}_p$. We apply ℓ_2-normalization to both $\boldsymbol{v}$ and $\boldsymbol{c}$ to have $(\boldsymbol{v} - \boldsymbol{c})^2 = 2 - 2\boldsymbol{v} \cdot \boldsymbol{c}$. Then combining this with Eqs. (2), (3), (6), (7), (8), and (9), we can get the maximum log-likelihood estimation with:

$$\begin{aligned} \theta^* &= \arg\min_{\theta} \sum_{i=1}^{N} -\log \frac{\exp\left(\boldsymbol{v}_i \cdot \boldsymbol{c}_p / \tau_p\right)}{\sum_{j=1}^{K} \exp\left(\boldsymbol{v}_i \cdot \boldsymbol{c}_j / \tau_j\right)} \\ \Longleftrightarrow \theta^* &= \arg\min_{\theta} \sum_{k=1}^{K}\sum_{i=1}^{N_k} -\log \frac{\exp\left(\boldsymbol{v}_i^k \cdot \boldsymbol{c}_k / \tau_k\right)}{\sum_{j=1}^{K} \exp\left(\boldsymbol{v}_i^k \cdot \boldsymbol{c}_j / \tau_j\right)}, \end{aligned} \tag{10}$$

where $\boldsymbol{v}_i^k$ denotes the representation of i^{th} sample (*i.e.*, skeleton sequence) belonging to the k^{th} prototype $\boldsymbol{c}_k$, N_k is the number of samples in the k^{th} cluster, and τ is related to the distribution of features around different prototypes.

Assumption 4.2. ***Maximum Homogeneous Similarity.*** The homogeneous instances, which are defined as instances within the same cluster, should share higher inherent similarity than heterogeneous instances between different clusters. In other words, the prototype of each cluster can represent a unique skeleton concept or attribute of a certain identity, and the same cluster's instances possess the homogeneity of features corresponding to this prototype [131]. According to Assumption 4.1, it can be equivalent to the objective that each instance should be maximally similar to the corresponding prototype and be minimally similar to other prototypes. In the M-step, we maximize the probability that each instance belongs to its unique prototype (see Eq. (9)) based on this assumption. The equivalent formulation of this objective in Eq. (10) after applying feature ℓ_2-normalization can be further interpreted as to maximize the dot-product based similarity between instances and their prototypes while maximizing the dissimilarity to other prototypes.

Relations to Existing Contrastive Losses: (1) The InfoNCE loss [132] used or re-formulated in MoCo [94] and SimCLR [91] can be interpreted as special cases of the maximum log-likelihood estimation in Eq. (10), where the prototype $\boldsymbol{c}_p$ for a feature $\boldsymbol{v}_i$ is replaced by the augmented feature $\boldsymbol{v'}_i$ generated from different views of augmentation of the same instance (*i.e.*, $\boldsymbol{c}_p = \boldsymbol{v'}_i$) and τ is fixed as a temperature for contrastive learning. **(2)** The ProtoNCE loss used in PCL [88] has a similar form as Eq. (10), where τ is estimated with the assumption that the distribution of feature representations around each prototype varies in different clusters. However, PCL estimates the feature distribution under the Euclidean distance metric used in the k-means clustering. Such estimation could be inapplicable (*e.g.*, can not be generalized) to models that employ different clustering algorithms (*e.g.*, density-based DBSCAN [6]) or/and different distance metrics (*e.g.*, Jaccard metric), thus failing to getting satisfactory performance in practice [5].

In our work, we adopt a generic form following InfoNCE loss, *i.e.*, setting a global temperature τ, for the proposed SimMC framework. By assuming a uniform feature distribution around each instance (*i.e.*, $\tau = \tau_k = \tau_j$), we select an appropriate τ to encourage the framework to learn representations with higher global uniformity,

which could improve the quality of contrastive representation learning as proved in [86, 130]. This is consistent with the results shown in Sec. 4.4.5.

In the proposed framework, we sample q subsequences of length $(f - x)$ by applying x random masks to each input skeleton sequence. Specifically, we generate q sets of subsequences for the whole training set by randomly sampling one subsequence for each training sequence at each round of q sampling rounds, which are encoded into corresponding skeleton instance sets by the embedding/encoder function $f_\theta(\cdot)$. Then we independently perform clustering on each instance set to obtain corresponding skeleton prototypes. The random subsequence sampling and multiple individual clustering encourage a more stable probability estimation of skeleton prototypes and facilitate mining more valuable intra-sequence skeleton features and high-level semantics (*e.g.*, class-related semantics) within skeleton sequences. We can formulate the proposed masked prototype contrastive (MPC) loss based on Eq. (10) as:

$$\mathcal{L}_{\text{MPC}} = \frac{1}{N} \sum_{i=1}^{q} \sum_{k=1}^{K_i} \sum_{j=1}^{N_k} -\log \frac{\exp\left(\boldsymbol{v}_{(i),j}^{k} \cdot \boldsymbol{c}_{(i)}^{k} / \tau\right)}{\sum_{u=1}^{K_i} \exp\left(\boldsymbol{v}_{(i),j}^{k} \cdot \boldsymbol{c}_{(i)}^{u} / \tau\right)}, \tag{11}$$

where N represents the number of all training instances, K_i denotes the number of skeleton prototypes generated from the i^{th} set of subsequence representations, N_k is the number of skeleton instances belonging to the k^{th} prototype $\boldsymbol{c}_{(i)}^{k}$ (equivalent to $\boldsymbol{p}_{(i)}^{k}$ in Sec. 4.2.3), and τ represents the global temperature for contrastive learning. When we utilize the original sequences without sampled subsequences for skeleton prototype learning, *i.e.*, $q = 1$ and $x = 0$ (see Eq. (3.2) and (3.3) in Sec. 4.2.3), the objective of Eq. (11) has exactly the same form as Eq. (10), which is defined as naïve prototype contrastive learning (denoted as NPC) in our work. Therefore, the proposed MPC scheme could be viewed as performing finer prototype contrastive learning with different subsequences (*i.e.*, NPC using original sequences is a special case of MPC), and allows the model to jointly attend to key skeleton patterns from different representation subspaces of the original sequences, which facilitates learning more discriminative skeleton representations for person re-ID, as demonstrated in our work.

4.5.1.1 Convergence Proof

We provide the proof for the convergence of the proposed MPC under modeling the maximum log-likelihood estimation (see Eq. (10)). Recall Eqs. (2) and (3) and let

$$\begin{aligned}\ell(\theta) = \sum_{i=1}^{N} \log p\left(\boldsymbol{x}_i; \theta\right) &= \sum_{i=1}^{N} \log \sum_{j=1}^{K} p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right) \\ &= \sum_{i=1}^{N} \log \sum_{j=1}^{K} Q\left(\boldsymbol{c}_j\right) \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right)}{Q\left(\boldsymbol{c}_j\right)} \\ &\geq \sum_{i=1}^{N} \sum_{j=1}^{K} Q\left(\boldsymbol{c}_j\right) \log \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right)}{Q\left(\boldsymbol{c}_j\right)} \end{aligned} \tag{12}$$

The above inequality holds with equality when $Q\left(\boldsymbol{c}_j\right) = p\left(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta\right)$ is a constant (see Eq. (4)).

In the t^{th} E-step, we have estimated the constant value $Q^{(t)}\left(\boldsymbol{c}_j\right) = p\left(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta^{(t)}\right)$. Then we have:

$$\ell\left(\theta^{(t)}\right) = \sum_{i=1}^{N} \sum_{j=1}^{K} Q^{(t)}\left(\boldsymbol{c}_j\right) \log \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta^{(t)}\right)}{Q^{(t)}\left(\boldsymbol{c}_j\right)} \tag{13}$$

For the t^{th} M-step, we fix $Q^{(t)}\left(\boldsymbol{c}_j\right) = p\left(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta^{(t)}\right)$ and train model parameters θ to maximize Eq. (13). In this way, we can always have:

$$\begin{aligned}\ell\left(\theta^{(t+1)}\right) &\geq \sum_{i=1}^{N} \sum_{j=1}^{K} Q^{(t)}\left(\boldsymbol{c}_j\right) \log \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta^{(t+1)}\right)}{Q^{(t)}\left(\boldsymbol{c}_j\right)} \\ &\geq \sum_{i=1}^{N} \sum_{j=1}^{K} Q^{(t)}\left(\boldsymbol{c}_j\right) \log \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta^{(t)}\right)}{Q^{(t)}\left(\boldsymbol{c}_j\right)} = \ell\left(\theta^{(t)}\right) \end{aligned} \tag{14}$$

The above result that $\ell\left(\theta^{(t)}\right)$ monotonically increases with more iterations suggests the convergence of the algorithm.

4.5.2 MIC Modeled as Expectation-Maximization Algorithm

The proposed masked intra-sequence contrastive learning (MIC) can also be modeled as an Expectation-Maximization (EM) like algorithm, which is similar to the EM formulation in Sec. 4.5.1. Specifically, the MIC implicitly involves two sets of variables, and solves two underlying sub-problems. The MIC loss function can be formulated as:

$$\mathcal{L}_{\text{MIC}}(\theta, \eta) = \mathbb{E}_{\boldsymbol{x}, \mathcal{M}} \left[\| f_\theta(\mathcal{M}(\boldsymbol{x})) - \eta_{\boldsymbol{x}} \|_2^2 \right], \tag{15}$$

where $f_\theta(\cdot)$ is the embedding/encoder function parameterized by θ. $\mathcal{M}(\cdot)$ is a random masking function, which could be viewed as an augmentation strategy to produce augmented instances (*i.e.*, subsequences) of the same skeleton sequence. For convenience, we simplify $\mathcal{M}(\cdot)$ with $\boldsymbol{x}$ as the only input. $\boldsymbol{x}$ is the input skeleton sequence. η denotes the set of variables related to representations of samples, and the subscript $_{\boldsymbol{x}}$ means using the index of sample to access a sub-vector of η. Intuitively, $\eta_{\boldsymbol{x}}$ can be interpreted as the representation of the skeleton sequence $\boldsymbol{x}$. The expectation $\mathbb{E}[\cdot]$ is over the distribution of skeleton sequences and mask-based augmentation. The mean square error $\| \cdot \|_2^2$ is equivalent to the cosine similarity as the vectors are all ℓ_2-normalized.

The objective of MIC is to minimize the mean square loss between the encoded representations of augmented skeleton instances and the representation of the skeleton instance. For the ease of analysis, we model MIC as an optimization problem by:

$$\min_{\theta, \eta} \mathcal{L}(\theta, \eta) \tag{16}$$

Analogous to the masked prototype contrastive learning (see Sec. 4.5.1.1), the problem in formula (16) can be solved by an alternating algorithm. Formally, we can alternate between solving these two sub-problems:

$$\eta^{(t)} \leftarrow \arg\min_{\eta} \mathcal{L}\left(\theta^{(t)}, \eta\right), \tag{17}$$

$$\theta^{(t+1)} \leftarrow \arg\min_{\theta} \mathcal{L}\left(\theta, \eta^{(t)}\right), \tag{18}$$

where t is the index of alternation and $\leftarrow$ denotes assigning.

In the t^{th} E-step (see Eq. (17)), the sub-problem is to minimize $\mathbb{E}_{\mathcal{M}}\left[\|f_{\theta^{(t)}}(\mathcal{M}(\boldsymbol{x})) - \eta_{\boldsymbol{x}}\|_2^2\right]$ for each skeleton sequence $\boldsymbol{x}$ given the fixed $\theta^{(t)}$. Recalling the nature of the mean square error, it can be solved by:

$$\eta_{\boldsymbol{x}}^{(t)} \leftarrow \mathbb{E}_{\mathcal{M}}\left[f_{\theta^{(t)}}(\mathcal{M}(\boldsymbol{x}))\right], \tag{19}$$

where $\eta_{\boldsymbol{x}}$ is assigned with the average representation of $\boldsymbol{x}$ over the distribution of mask-based augmentation.

For the t^{th} M-step (see Eq. (18)), the $\eta^{(t)}$ is naturally fixed as the gradient does not back-propagate to $\eta^{(t)}$, which is a constant in this sub-problem. We train model parameters θ to maximize the inherent similarity, *i.e.*, minimizing the mean square loss, between $f_\theta(\mathcal{M}(\boldsymbol{x}))$ and $\eta_{\boldsymbol{x}}^{(t)}$ to get $\theta^{(t+1)}$.

4.5.2.1 Hypothesis for Effectiveness of Siamese Architectures

The proposed MIC can be approximated by one-step alternation between Eqs. (17) and (18). In particular, we approximate Eq. (19) by sampling the mask-based augmentation only *once*, denoted as $\mathcal{M}'$, and ignoring $\mathbb{E}_{\mathcal{M}}[\,\cdot\,]$:

$$\eta_{\boldsymbol{x}}^{(t)} \leftarrow f_{\theta^{(t)}}\left(\mathcal{M}'(\boldsymbol{x})\right) \tag{20}$$

By inserting it into the sub-problem (see Eq. (18)) and combining Eq. (15), we can have:

$$\theta^{(t+1)} \leftarrow \arg\min_{\theta} \mathbb{E}_{x,\mathcal{M}}\left[\|f_\theta(\mathcal{M}(\boldsymbol{x})) - f_{\theta^{(t)}}\left(\mathcal{M}'(\boldsymbol{x})\right)\|_2^2\right], \tag{21}$$

where $\theta^{(t)}$ is a constant in this sub-problem, and $\mathcal{M}'$ implies another random view generated by the proposed mask-based augmentation (*i.e.*, random subsequence sampling). The formulation in Eq. (21) exhibits a Siamese architecture naturally with stop-gradient applied, which leads to the hypothesis that the Siamese network matches the proposed EM modeling and is efficient for the optimization. This hypothesis is also theoretically and empirically proved in SimSiam [95].

4.5.2.2 Hypothesis for Effectiveness of Fully-Connected Layer

The predictor head attached on one side of the Siamese architecture is a commonly-adopted architecture in contrastive learning models [91, 95]. In our framework, we adopt a fully-connected (FC) layer (denoted as $\mathcal{F}_c(\cdot)$) rather than an MLP network as the predictor head. Here we hypothesize that $\mathcal{F}_c$ is beneficial for the approximation of Eq. (20). Specifically, the predictor head $\mathcal{F}_c(\cdot)$ is expected to minimize $\mathbb{E}_{\boldsymbol{v}}\left[\|\mathcal{F}_c(\boldsymbol{v}_1)-\boldsymbol{v}_2\|_2^2\right]$, where $\boldsymbol{v}_1=f_\theta(\mathcal{M}(\boldsymbol{x}_1)), \boldsymbol{v}_2=f_\theta(\mathcal{M}(\boldsymbol{x}_2))$. Therefore, the optimal solution to $\mathcal{F}_c$ should satisfy: $\mathcal{F}_c(\boldsymbol{v}_1)=\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{v}_2]=\mathbb{E}_{\mathcal{M}}[f_\theta(\mathcal{M}(\boldsymbol{x}))]$ for any instance of skeleton sequence $\boldsymbol{x}$. This term is similar to the one in Eq. (19). Since the expectation $\mathbb{E}_{\mathcal{M}}$ is ignored in the approximation of Eq. (20), the hypothesis is that the application of $\mathcal{F}_c(\cdot)$ could fill this gap. Considering that directly computing the expectation $\mathbb{E}_{\mathcal{M}}$ is intractable and the masking function $\mathcal{M}$ that samples random subsequences can be viewed as to linearly combine different skeletons of the same sequence, $\mathbb{E}_{\mathcal{M}}[f_\theta(\mathcal{M}(\boldsymbol{x}))]$ is assumed to be a linear transformation of skeleton features in the latent space, thus we exploit a linear FC layer $\mathcal{F}_c(\cdot)$ to learn to predict the expectation, under the condition that the sampling of $\mathcal{M}$ is implicitly distributed across multiple epochs. Theoretically, an MLP network is also feasible to approximate such expectation. In practice, using the FC layer or MLP can achieve similar performance on different datasets (see Sec. 4.4.4), and the FC layer is adopted to enjoy smaller parameter size and lower computational complexity.

4.6 Summary

In this chapter, we propose a simple masked contrastive learning (SimMC) framework to efficiently learn representations of unlabeled skeleton sequences for unsupervised person re-ID. A novel masked prototype contrastive learning (MPC) scheme is devised to cluster the most typical skeleton features of subsequences randomly masked from original sequences, so as to contrast their inherent similarity to learn a discriminative skeleton representation from unlabeled skeletons. To fully exploit inherent relationships between subsequences, we propose a masked intra-sequence contrastive learning (MIC) to learn their similarity and pattern consistency within the sequence for more effective skeleton representations. Our

framework outperforms most state-of-the-art skeleton-based methods and also enjoys high scalability and efficiency to be applied to different models and scenes. Furthermore, we theoretically model our framework as expectation-maximization (EM) solutions to prove its effectiveness and convergence.

Chapter 5

Skeleton-Based Person Re-ID with Multi-Level Body Modeling

5.1 Introduction

Traditional skeleton-based methods [2, 23, 24, 34] typically extract hand-crafted features like skeleton descriptors in terms of pre-defined anthropometric and gait attributes of body, and recent mainstream methods [3, 8, 25] resort to deep neural networks (DNNs) to perform skeleton representation learning. Despite the great efforts, they usually require manually-annotated skeleton data to train or fine-tune models, which is labor-expensive and could reduce their general applicability in practice. Therefore, how to fully exploit *unlabeled* skeleton data to learn general effective skeleton representations for person re-ID is a key challenge in this area.

Although a few works such as the proposed SimMC framework [5] (see Chapter 4) have explored unsupervised skeleton learning, they typically rely on clustering and contrastive learning in a *single* feature space, where the inherent randomness of feature initialization or clustering [88] could induce unstable feature distribution estimation and impede the skeleton representation learning (*e.g.*, prototype learning) (detailed in Sec. 5.2.1). Another crucial shortcoming of existing methods is that they typically learn skeleton features from a single level (*e.g.*, body joint

This chapter has been published as: Haocong Rao, Cyril Leung, and Chunyan Miao, "Hierarchical Skeleton Meta-Prototype Contrastive Learning with Hard Skeleton Mining for Unsupervised Person Re-Identification," *International Journal of Computer Vision (IJCV)*, 2023 [133]. DOI: 10.1007/s11263-023-01864-0.

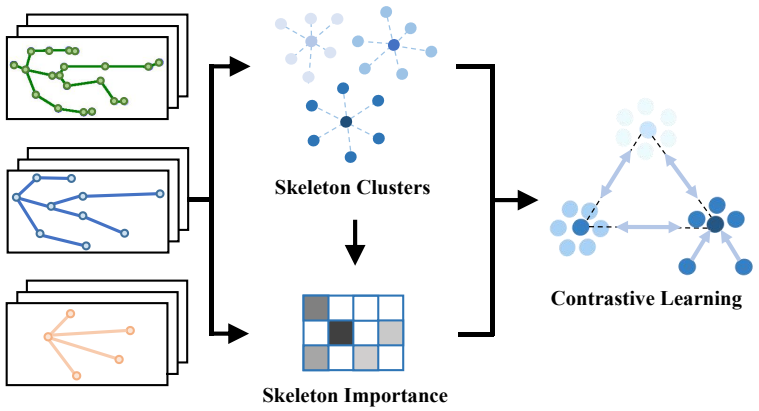


FIGURE 5.1: Overview of Hi-MPC approach: It hierarchically clusters skeleton representations to infer their inherent importance, and contrasts the key clustered features with the most typical ones to learn effective representations for person re-ID.

level [3, 4]) and assume that each skeleton is equally important in representing the patterns of a person [4, 5]. This intrinsically limits their ability to exploit key features of more informative skeletons. For instance, there usually exist skeletons that are either harder to be recognized as the same identity (*i.e.*, large intra-class variations) or easier to be misidentified among different individuals (*i.e.*, small inter-class variations), both of which should be given greater attention for learning more effective skeleton representations. In summary, previous methods typically lack the ability to automatically learn effective skeleton representations from *unlabeled* skeleton data (corresponding to the first challenge in Sec. 1.2), and cannot mine valuable hierarchical (*i.e.*, different level) body and motion features from skeleton data (corresponding to the second challenge in Sec. 1.2).

To solve the aforementioned challenges, we propose a generic Hierarchical skeleton Meta-Prototype Contrastive learning (Hi-MPC) approach with a Hard Skeleton Mining (HSM) mechanism[1]. As shown in Fig. 5.1, the proposed approach exploits *unlabeled* hierarchical representations of key informative skeletons (*i.e.*, with higher importance inferred) to contrast and learn the most typical skeleton features for person re-ID. Firstly, we construct a hierarchical representation for each skeleton with coarse-to-fine body partitions, which enables the model to explore body structures and pattern information from different levels. Secondly, a Hierarchical skeleton Meta-Prototype Contrastive learning (Hi-MPC) approach is devised to cluster and contrast the most representative skeleton features (defined as *"prototypes"*)

[1]Our codes are publicly available at https://github.com/Kali-Hac/Hi-MPC.

from *different-level* skeleton representations (defined as *"instances"*). To encourage learning more consistent and representative skeleton prototypes, we propose to transform original instances into multiple *homogeneous meta-instances*, and maximize their inherent similarity to corresponding *meta-prototypes* while maximizing their dissimilarity to others, so as to capture discriminative skeleton features and class-related semantics (*e.g.*, intra-class similarity) from various levels of unlabeled skeletons. Thirdly, considering that different skeletons usually possess varying informative value, *e.g.*, some skeletons are more difficult to be classified to the correct identity (defined as *"hard skeletons"*) but can provide more informative clues for model learning [39], we *for the first time* devise a Hard Skeleton Mining (HSM) mechanism to adaptively infer the importance of each skeleton in learning hard and easily-confused patterns. In this way, HSM enables our model to mine and focus on hard skeletons in Hi-MPC to encourage more effective skeleton representation learning. Lastly, we propose to construct the novel Multi-level Skeleton Meta-Representation (MSMR) that combines skeleton features learned from different levels as the final representation for person re-ID. Extensive experiments on five public benchmarks demonstrate that our approach outperforms most state-of-the-art methods on person re-ID tasks. We further show that our method is generally effective in multi-view and RGB-based scenarios with estimated skeletons. As a byproduct of our approach, we also reveal the feasibility of exploiting more concise and abstract skeleton representations to perform person re-ID.

With this chapter, we make the following contributions:

- We devise hierarchical representations of 3D skeletons and propose a novel hierarchical skeleton meta-prototype contrastive learning approach with a hard skeleton mining mechanism to learn effective representations from *unlabeled* skeleton sequences for person re-ID.
- We propose the Hierarchical skeleton Meta-Prototype Contrastive learning (Hi-MPC) that hierarchically contrasts representative features and inherent similarity of different-level skeleton representations to learn discriminative features and high-level semantics for person re-ID.
- We devise a Hard Skeleton Mining (HSM) mechanism to adaptively infer informative importance of skeletons within each sequence to encourage learning more effective skeleton representations from harder skeletons.

- We empirically validate the effectiveness of each level skeleton representation learned from the proposed approach, and combine them to construct the novel Multi-level Skeleton Meta-Representation (MSMR) for person re-ID.
- Extensive experiments on five public benchmarks demonstrate that our approach outperforms most state-of-the-art methods on person re-ID tasks. We further show that our method is generally effective in multi-view and RGB-based scenarios with estimated skeletons.

5.2 The Proposed Hi-MPC Approach

The goal of our approach is to perform *unsupervised* person re-identification with unlabeled 3D skeleton sequences. Formally, we denote a sequence of 3D skeletons as $\boldsymbol{S}_{1:F} = (\boldsymbol{S}_1, \cdots, \boldsymbol{S}_F) \in \mathbb{R}^{F \times K}$, where $\boldsymbol{S}_i \in \mathbb{R}^K$ is the i^{th} skeleton with 3D positions of J body joints and $K = 3 \times J$. The training set $\Phi^{\mathcal{T}} = \left\{\boldsymbol{S}_{1:F}^{\mathcal{T},i}\right\}_{i=1}^{N_1}$, probe set $\Phi^{\mathcal{P}} = \left\{\boldsymbol{S}_{1:F}^{\mathcal{P},i}\right\}_{i=1}^{N_2}$, and gallery set $\Phi^{\mathcal{G}} = \left\{\boldsymbol{S}_{1:F}^{\mathcal{G},i}\right\}_{i=1}^{N_3}$ contain N_1, N_2, and N_3 skeleton sequences of different persons in varying views or occasions. Each skeleton sequence $\boldsymbol{S}_{1:F}$ corresponds to a unique identity y $\in \{1, \cdots, I\}$ where I is the number of different identities. Our approach aims at learning to encode $\Phi^{\mathcal{P}}$ and $\Phi^{\mathcal{G}}$ into effective skeleton representations $\{\boldsymbol{V}_i^{\mathcal{P}}\}_{i=1}^{N_2}$ and $\{\boldsymbol{V}_j^{\mathcal{G}}\}_{j=1}^{N_3}$ *without using any label*, such that the representation $\boldsymbol{V}_i^{\mathcal{P}}$ in the probe set can match the representation $\boldsymbol{V}_j^{\mathcal{G}}$ of the same identity in the gallery set.

The overview of our approach is presented in Fig. 5.2. Firstly, each skeleton sequence $\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_F$ is hierarchically represented at joint-level (top), component-level (middle), and limb-level (bottom). Secondly, we perform temporal average pooling (TAP) on the encoded skeleton representations of each level to generate skeleton instances, and cluster them to find prototypes, which are then transformed into meta-instances and meta-prototypes in different contrastive subspaces. Lastly, a hard skeleton mining mechanism (illustrated in Fig. 5.3) is employed to infer informative importance of skeletons within each sequence, which is integrated into the contrastive loss $\mathcal{L}_{\text{Hi-MPC}^h}$ to enhance the similarity of key meta-instances belonging to the same meta-prototype while maximizing their dissimilarity to other meta-prototypes. The meta-instances learned from different levels are combined to

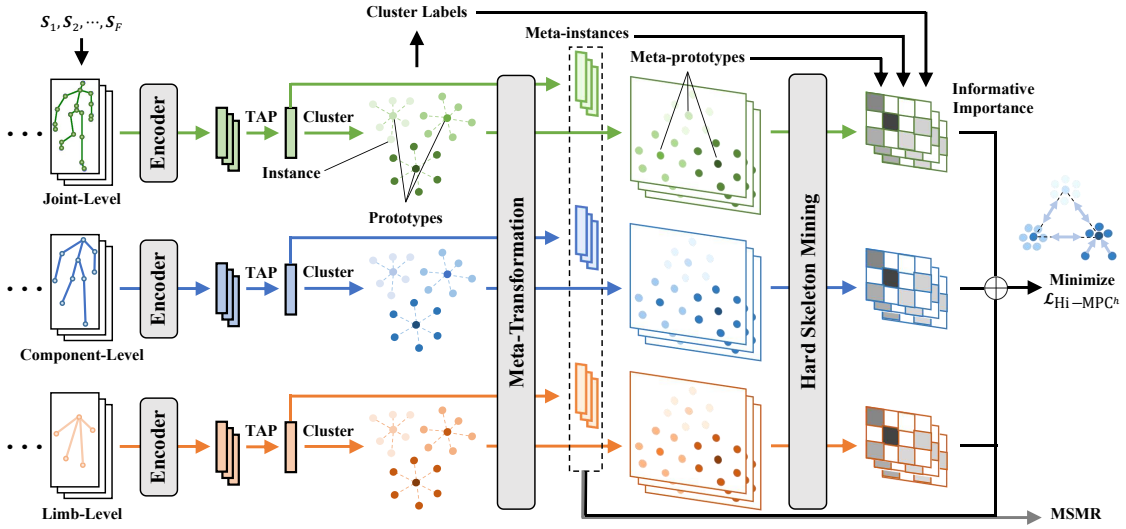


FIGURE 5.2: Schematics of our approach with hierarchical skeleton representations, hierarchical skeleton meta-prototype contrastive learning, and hard skeleton mining mechanism (detailed in Fig. 5.3).

construct multi-level skeleton meta-representation (MSMR) for the person re-ID task. We detail each technical component below.

5.2.1 Hierarchical Skeleton Representations

The human body can be naturally modeled with several key functional regions at different levels (*e.g.*, joints, limbs) [7, 113], which could *hierarchically* characterize different anthropometric or kinetic features of the body. Inspired by this fact, we spatially divide each human skeleton into various non-overlapping partitions, each of which corresponds to a certain body part at the level of joints, body components (*e.g.*, hands) or limbs (*e.g.*, upper limbs). Then we generate the position of each body part by computing the centroid of body joints within the corresponding partition. As presented in Fig. 5.2, we build hierarchical skeleton representations for each skeleton $\boldsymbol{S}$ from three levels, namely *joint-level*, *component-level*, and *limb-level* skeleton representations, which can correspondingly contain low, middle, and high level body structures and pattern information of a skeleton. Formally, the l^{th} level representation $\hat{\boldsymbol{S}}^l \in \mathbb{R}^{3\times n_l}$ ($l \in \{1,2,3\}$) consists of 3D positions of n_l body partitions, where $n_1 = J$, $n_2 = 10$, $n_3 = 5$ correspond to joint-level ($\hat{\boldsymbol{S}}^1$), component-level ($\hat{\boldsymbol{S}}^2$), limb-level skeleton representations ($\hat{\boldsymbol{S}}^3$), respectively. We denote the hierarchical representations of each input skeleton sequence $\boldsymbol{S}_{1:F}$ as $\hat{\boldsymbol{S}}^1_{1:F} \in \mathbb{R}^{F\times 3n_1}$, $\hat{\boldsymbol{S}}^2_{1:F} \in \mathbb{R}^{F\times 3n_2}$, and $\hat{\boldsymbol{S}}^3_{1:F} \in \mathbb{R}^{F\times 3n_3}$.

5.2.2 Hierarchical Skeleton Meta-Prototype Contrastive Learning

As each pedestrian's skeletons typically possess unique identity-associated features in terms of anthropometric attributes (*e.g.*, body part lengths) and walking patterns [37], it is desirable to exploit the *most typical skeleton features ("prototypes")* from skeleton sequences *("instances")* to differentiate a given person from others. A straightforward solution is to find the skeleton prototypes by clustering skeleton instances for direct prototype-instance contrastive learning in a *single* feature space [5]. However, the inherent randomness of feature initialization or clustering [88] could induce unstable prototype estimation and inconsistent relational distributions (*e.g.*, prototype-instance relations) when representation spaces vary. Based on the assumption that the global distribution of prototypes should be consistent with the distribution of cluster centroids (referred to as *"prototype-cluster consistency"*, see Assumption 5.1), we propose to construct different contrastive subspaces that inherit from the original feature space of prototypes to enhance contrastive learning. In particular, we perform prototype-instance contrasting in each *individual* contrastive subspace, which are combined based on the prototype-cluster consistency to encourage more robust probability estimation of prototypes and more consistent contrastive learning. To this end, we devise the ***hierarchical skeleton meta-prototype contrastive learning (Hi-MPC)*** to homogeneously transform original prototypes and instances into ***meta-prototypes*** and ***meta-instances*** at each skeleton level, and contrast their inherent similarity in different transformed contrastive subspaces to *jointly* learn representative discriminative skeleton features for person re-ID.

Given the l^{th} level skeleton representations $\hat{\boldsymbol{S}}_1^l, \cdots, \hat{\boldsymbol{S}}_F^l$ of an input skeleton sequence, we first encode them and apply temporal average pooling (TAP) to obtain a sequence-level skeleton representation, *i.e.*, instance (shown in Fig. 5.2) as:

$$\boldsymbol{v}^{l,(i)} = \frac{1}{F}\sum_{j=1}^{F}\psi^l\left(\hat{\boldsymbol{S}}_j^{l,(i)}\right) = \frac{1}{F}\sum_{j=1}^{F}\boldsymbol{z}_j^{l,(i)}, \tag{5.1}$$

where $\psi^l(\cdot)$ is the l^{th} level encoder built by a multi-layer perceptron (MLP) network with one hidden layer, $\boldsymbol{z}_j^{l,(i)} \in \mathbb{R}^{h_1}$ denotes the encoded features of the l^{th} level representation of j^{th} skeleton in the i^{th} training skeleton sequence, and $\boldsymbol{v}^{l,(i)} \in$

$\mathbb{R}^{h_1}$ denotes the encoded l^{th} level representation of i^{th} training skeleton sequence $\hat{\boldsymbol{S}}_{1:F}^{l,(i)}, i \in \{1, \cdots, N_1\}$. Here we adopt TAP to *average* the temporal dynamics of all skeletons to represent the features of a sequence [5]. It is worth noting that TAP also keeps the consistency of feature dimensions between skeleton-level and sequence-level representations. This allows us to directly compute their inherent similarity by dot products without extra dimension transformation (see Sec. 5.2.3).

Then, to mine the original skeleton prototypes, we exploit the encoded sequence-level representations $\mathbb{V}^l = \{\boldsymbol{v}^{l,(1)}, \cdots, \boldsymbol{v}^{l,(N_1)}\}$ as skeleton instances, and leverage the DBSCAN algorithm [6] to cluster instances of similar features and semantics. As shown in Fig. 5.2, we generate clusters as $\widehat{\mathbb{V}}_c^l = \{\boldsymbol{v}_{c,k}^l\}_{k=1}^{n_c}$, $c \in \{1, \cdots, C\}$, where C denotes the number of different clusters, *i.e.*, pseudo classes, and the c^{th} cluster $\widehat{\mathbb{V}}_c^l$ contains n_c instances. Note that we perform clustering *individually* on each level of hierarchical skeleton representations to better capture different level semantics and retain coarse-to-fine skeleton features. The instance features of the same cluster are averaged as the corresponding skeleton prototype with:

$$\boldsymbol{p}_c^l = \frac{1}{n_c} \sum_{k=1}^{n_c} \boldsymbol{v}_{c,k}^l \ , \tag{5.2}$$

where $\boldsymbol{p}_c^l \in \mathbb{R}^{h_1}$ denotes the original skeleton prototype of the c^{th} cluster $\widehat{\mathbb{V}}_c^l$ generated from the l^{th} level skeleton instances. Given the original prototypes and instances, our model converts them into *meta-prototypes* and *meta-instances* with multiple meta-transformation heads by:

$$(\hat{\boldsymbol{v}}_{c,k}^l)^m = \mathbf{H}_1^{l,m} \boldsymbol{v}_{c,k}^l \ , \tag{5.3}$$

$$(\hat{\boldsymbol{p}}_c^l)^m = \mathbf{H}_2^{l,m} \boldsymbol{p}_c^l \ , \tag{5.4}$$

where $(\hat{\boldsymbol{v}}_{c,k}^l)^m, (\hat{\boldsymbol{p}}_c^l)^m \in \mathbb{R}^{h_2}$ denote the m^{th} meta-instance and meta-prototype transformed from $\boldsymbol{v}_{c,k}^l$ and $\boldsymbol{p}_c^l$. Here $\mathbf{H}_1^{l,m}, \mathbf{H}_2^{l,m} \in \mathbb{R}^{h_2 \times h_1}$ are corresponding learnable weight matrices of the m^{th} transformation head. $m \in \{1, \cdots, M\}$ and M denotes the number of different transformation heads. Considering that both original instances and prototypes come from the same domain, *i.e.*, being represented in the *homogeneous* feature space of the same dimension [134], it is natural to employ homogeneous feature mapping for each pair of heads (defined as

"meta-transformation heads") with $\mathbf{H}_1^{l,m} = \mathbf{H}_2^{l,m}$ and $h = h_1 = h_2$. The meta-transformation heads map both instances and prototypes into the *same* m^{th} new feature space, which can be viewed as the m^{th} subspace *linearly* transformed from the original contrastive feature space, to generate homogeneous meta-instances and meta-prototypes. It should be noted that we do NOT use *heterogeneous* feature mapping (*i.e.*, $\mathbf{H}_1^{l,m} \neq \mathbf{H}_2^{l,m}$), as it separately maps instances and prototypes into two different feature subspaces with domain shifts [135] and degrades the model performance (see Sec. 5.4.7.5).

To jointly focus on representative skeleton features of all meta-prototypes and capture different-level skeleton semantics (*e.g.*, class-related patterns) from diverse contrastive feature subspaces, we propose the Hi-MPC loss below:

$$\mathcal{L}_{\text{Hi-MPC}} = \sum_{l=1}^{3}\sum_{i=1}^{I_l}\sum_{m=1}^{M} -\log \frac{\exp\left((\hat{\boldsymbol{v}}^{l,(i)})^m \cdot (\hat{\boldsymbol{p}}_+^l)^m/\tau\right)}{\sum_{c=1}^{C}\exp\left((\hat{\boldsymbol{v}}^{l,(i)})^m \cdot (\hat{\boldsymbol{p}}_c^l)^m/\tau\right)}, \tag{5.5}$$

where I_l denotes the number of instances in all clusters generated from the l^{th} level skeleton representations, $(\hat{\boldsymbol{v}}^{l,(i)})^m$ denotes the m^{th} transformed meta-instance of i^{th} instance belonging to its corresponding meta-prototype $(\hat{\boldsymbol{p}}_+^l)^m$, $(\hat{\boldsymbol{p}}_c^l)^m$ is the meta-prototype of the c^{th} cluster at the l^{th} level, and τ is the temperature for contrastive learning. We set $\tau = \sqrt{h}$ to scale the dot products to improve the stability of contrastive learning [108]. Note that $\mathcal{L}_{\text{Hi-MPC}}$ is averaged over all meta-instances for training. The Hi-MPC approach combining both hierarchical skeleton clustering (see Eqs. (5.1) and (5.2)) and multiple meta-transformations (see Eqs. (5.3) and (5.4)) enables our model to perform a coarse-to-fine skeleton prototype estimation and mine different-level skeleton semantics (*e.g.*, identity-specific semantics), and also encourages more consistent prototype learning by jointly attending to key meta-prototypes in different representation subspaces. However, Hi-MPC only considers sequence-level skeleton representations, *i.e.*, instances with averaged skeleton features (see Eq. (5.1)), and cannot fully exploit key skeletons with higher importance in each sequence for contrastive learning, which motivates us to propose the hard skeleton mining mechanism below.

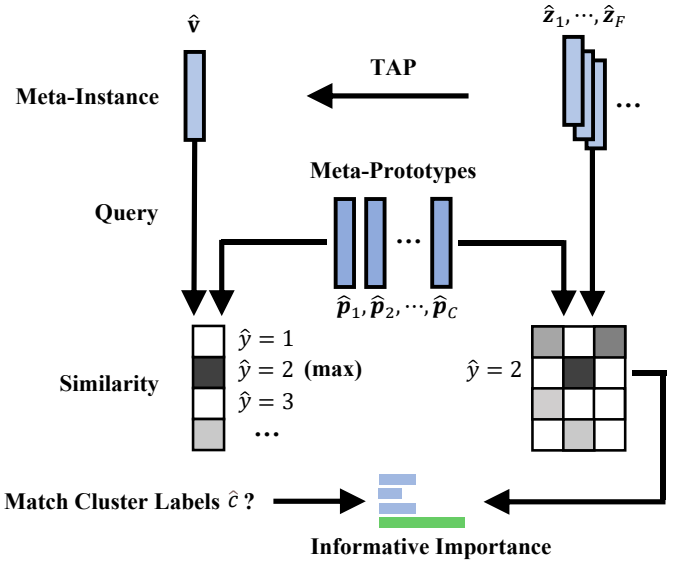


FIGURE 5.3: Overview and computation flow of hard skeleton mining (HSM) mechanism. The similarity between skeleton representations $\hat{\boldsymbol{z}}_1, \cdots, \hat{\boldsymbol{z}}_F$ and the meta-prototype ($\hat{\mathrm{y}}$) predicted by the meta-instance $\hat{\boldsymbol{v}}$ is first queried. The informative importance is then inferred based on the true or false matching of the cluster label $\hat{c}$.

5.2.3 Hard Skeleton Mining Mechanism

Different skeletons within the same sequence could possess different importance (*referred to as "**informative importance**"*) in mining hard (*e.g.*, easily-confused) patterns of a person. In particular, similar-looking skeletons and patterns shared among different persons (*referred to as "hard negatives"*), or wildly different poses of the same person (*referred to as "hard positives"*), are typically harder to be distinguished while they can provide more informative clues for models to comprehend the full concept of "same person" [39]. To achieve this goal, we propose the ***hard skeleton mining (HSM) mechanism*** to encourage the model to focus on skeleton representations with higher informative importance from hard negatives and hard positives in Hi-MPC.

As shown in Fig. 5.3, given the skeleton sequence representation $\hat{\boldsymbol{v}}$ that belongs to the $\hat{c}^{th}$ cluster, we first predict its label by querying the dot product based similarity with meta-prototypes in the m^{th} contrastive feature subspace by:

$$\hat{\mathrm{y}}^m = \arg\max_i \left((\hat{\boldsymbol{v}})^m \cdot (\hat{\boldsymbol{p}}_i)^m \right). \tag{5.6}$$

In Eq. (5.6), $\hat{\mathrm{y}}^m \in \{1, \cdots, C\}$ is the predicted cluster label, $(\hat{\boldsymbol{v}})^m$ and $(\hat{\boldsymbol{p}}_i)^m$ denote the meta-instance and the i^{th} meta-prototype generated by the m^{th} meta-transformation head. It is worth noting that the *cluster label* (denoted as $\hat{c}$) generated by DBSCAN algorithm is adopted as the ground-truth label since the real label is not available. We use the label of the cluster centroid (*i.e.*, meta-prototype) that has the maximum similarity with $(\hat{\boldsymbol{v}})^m$ as the predicted cluster label of $(\hat{\boldsymbol{v}})^m$ in the m^{th} contrastive feature subspace. For convenience, in Eq. (5.6) we omit the superscripts of levels and use $\hat{\boldsymbol{S}}_{1:F}$, $(\hat{\boldsymbol{v}})^m$, and $(\hat{\boldsymbol{p}}_i)^m$ to denote the l^{th} level representation $\hat{\boldsymbol{S}}^l_{1:F}$, $(\hat{\boldsymbol{v}}^l)^m$, and $(\hat{\boldsymbol{p}}^l_i)^m$, while $\hat{\mathrm{y}}^m$ corresponds to the predicted cluster label of the meta-instance $(\hat{\boldsymbol{v}})^m$.

Algorithm 1 Main Algorithm of Hi-MPC with HSM

Input: Unlabeled training skeleton sequences $\Phi^{\mathcal{T}} = \left\{ \boldsymbol{S}^{\mathcal{T},(i)}_{1:F} \right\}_{i=1}^{N_1}$, initialized encoder function $\psi(\cdot)$, initialized M meta-transformation heads $\mathrm{Meta}^m(\cdot)$, temperature τ

Output: Encoder $\psi(\cdot)$, meta-transformation head $\mathrm{Meta}^m(\cdot)$

$\hat{\boldsymbol{S}}^{1,(i)}_{1:F}, \hat{\boldsymbol{S}}^{2,(i)}_{1:F}, \hat{\boldsymbol{S}}^{3,(i)}_{1:F} = \mathrm{Hier}\left(\boldsymbol{S}^{\mathcal{T},(i)}_{1:F}\right)$
▷ Hierarchical skeleton representations at joint/component/limb-level

repeat

$\boldsymbol{v}^{l,(i)} = \mathrm{TAP}\left(\psi(\hat{\boldsymbol{S}}^{l,(i)}_{1:F})\right) = \mathrm{TAP}\left(\boldsymbol{z}^{l,(i)}_1, \cdots, \boldsymbol{z}^{l,(i)}_F\right)$
▷ Encode hierarchical skeleton sequences into instances

$\{\widehat{\mathbb{V}}^l_c\}_{c=1}^{C} = \mathrm{DBSCAN}\left(\{\boldsymbol{v}^{l,(i)}\}_{i=1}^{N_1}\right)$ ▷ Find clusters and discard outliers

$\boldsymbol{p}^l_c = \mathrm{Proto}\left(\widehat{\mathbb{V}}^l_c\right)$ ▷ Generate skeleton prototypes with Eq.(5.2)

$\left((\hat{\boldsymbol{v}}^l_{c,k})^m, (\hat{\boldsymbol{p}}^l_c)^m\right) = \mathrm{Meta}^m\left(\boldsymbol{v}^l_{c,k}, \boldsymbol{p}^l_c\right)$
▷ Perform the m^{th} meta-transformation with Eq.(5.3), (5.4)

$\hat{\mathrm{y}}^m = \mathrm{Pred}\left((\hat{z}^l)^m\right) = \mathrm{Pred}\left((\hat{v}^l)^m\right)$
▷ Predict cluster label for $(\hat{v}^l)^m$ and its skeletons with Eq.(5.6)

Use predicted meta-prototype $(\hat{\boldsymbol{p}}_{\hat{\mathrm{y}}^m})^m$ to infer importance $\overline{\delta}\left((\hat{z}^l)^m\right)$ of each skeleton with Eq. (5.8)

$\mathcal{L}_{\text{Hi-MPC}^h}\left(\overline{\delta}\left((\hat{\boldsymbol{z}}^l_j)^m\right), (\hat{\boldsymbol{z}}^l_j)^m, \{(\hat{\boldsymbol{p}}^l_c)^m\}_{c=1}^{C}, \tau\right)$
▷ Compute importance-weighted contrastive loss with Eq.(5.9)

Update parameters of $\psi(\cdot)$ and $\mathrm{Meta}^m(\cdot)$ to minimize $\mathcal{L}_{\text{Hi-MPC}^h}$

until *MaxEpoch* or *MaxPatience*

Having obtained encoded features $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_F$ of F skeletons in $\hat{\boldsymbol{S}}_{1:F}$, where $\boldsymbol{z}_j = \psi\left(\hat{\boldsymbol{S}}_j\right)$ and $j \in \{1, \cdots, F\}$ (see Eq. (5.1)), we assign the cluster label $\hat{\mathrm{y}}^m$ predicted by their sequence-level representation $(\hat{\boldsymbol{v}})^m$ to each of them. As illustrated in Fig. 5.3, we compute the inherent similarity of each skeleton representation to the

predicted cluster using:

$$\delta((\hat{\boldsymbol{z}}_j)^m) = \frac{\exp\left((\hat{\boldsymbol{z}}_j)^m \cdot (\hat{\boldsymbol{p}}_{\hat{\mathrm{y}}^m})^m\right)}{\sum_{t=1}^{F} \exp\left((\hat{\boldsymbol{z}}_t)^m \cdot (\hat{\boldsymbol{p}}_{\hat{\mathrm{y}}^m})^m\right)}. \tag{5.7}$$

In Eq. (5.7), $\delta((\hat{\boldsymbol{z}}_j)^m) \in (0,1)$ represents the *normalized* similarity between the representation of j^{th} skeleton and the meta-prototype corresponding to the predicted $\hat{\mathrm{y}}^m$ cluster in the m^{th} contrastive subspace. For clarity and consistency, we use $(\hat{\boldsymbol{z}}_j)^m$ to represent the j^{th} skeleton representation transformed by the m^{th} head corresponding to Eq. (5.3). $\delta((\hat{\boldsymbol{z}}_j)^m)$ can be interpreted as the the degree of certainty that the j^{th} skeleton within the sequence is classified to $\hat{\mathrm{y}}^m$ in the m^{th} feature subspace, while higher certainty indicates that the skeleton is easier for learning to realize correct classification. Hence, the informative importance of each skeleton in the same sequence can be inferred by:

$$\overline{\delta}((\hat{\boldsymbol{z}}_j)^m) = \frac{\exp\left(\mathbb{I}(\hat{\mathrm{y}}^m, \hat{c})\left((\hat{\boldsymbol{z}}_j)^m \cdot (\hat{\boldsymbol{p}}_{\hat{\mathrm{y}}^m})^m\right)\right)}{\sum_{t=1}^{F} \exp\left(\mathbb{I}(\hat{\mathrm{y}}^m, \hat{c})\left((\hat{\boldsymbol{z}}_t)^m \cdot (\hat{\boldsymbol{p}}_{\hat{\mathrm{y}}^m})^m\right)\right)}, \tag{5.8}$$

where $\overline{\delta}((\hat{\boldsymbol{z}}_j)^m) \in (0,1)$ represents the informative importance of the j^{th} skeleton in the sequence $\hat{\boldsymbol{S}}_{1:F}$ when being represented in the m^{th} contrastive feature subspace, and $\mathbb{I}(\hat{\mathrm{y}}^m, \hat{c}) = -1$ if the predicted label $\hat{\mathrm{y}}^m$ of $(\hat{\boldsymbol{z}}_j)^m$ is $\hat{c}$ otherwise $\mathbb{I}(\hat{\mathrm{y}}^m, \hat{c}) = 1$. Intuitively, when the label prediction is true, *i.e.*, $\hat{\mathrm{y}}^m = \hat{c}$, the hardest positive is the skeleton with the *lowest* certainty $\delta(\cdot)$, which is more likely to contain diverse patterns of the same person and possesses higher informative importance, thus we have $\mathbb{I}(\hat{\mathrm{y}}^m, \hat{c}) = -1$ and $\overline{\delta}(\cdot) \propto \frac{1}{\delta(\cdot)}$. When the model fails to predict correctly, *i.e.*, $\hat{\mathrm{y}}^m \neq \hat{c}$, the hardest negative is the most certain skeleton being classified to the false label, while it contains more similar information that needs to be carefully distinguished. In this case, we have $\overline{\delta}(\cdot) \propto \delta(\cdot)$, which is naturally achieved with $\mathbb{I}(\hat{\mathrm{y}}^m, \hat{c}) = 1$. To facilitate coarse-to-fine pattern learning, the proposed HSM is performed on each level of skeleton hierarchical representations.

To fully exploit skeletons within each sequence and focus on harder skeletons with higher informative importance for Hi-MPC training, we integrate the skeleton importance into contrastive learning by proposing the Hi-MPC^h loss as follows:

$$\mathcal{L}_{\text{Hi-MPC}^h} = \sum_{l=1}^{3} \sum_{i=1}^{I_l} \sum_{j=1}^{F} \sum_{m=1}^{M} \overline{\delta}\left((\hat{\boldsymbol{z}}_j^{l,(i)})^m\right) \text{Softmax}_j\left((\hat{\boldsymbol{z}}_j^{l,(i)})^m \cdot (\hat{\boldsymbol{p}}_+^l)^m / \tau\right), \tag{5.9}$$

where $\text{Softmax}_j\left((\hat{\boldsymbol{z}}_j^{l,(i)})^m \cdot (\hat{\boldsymbol{p}}_+^l)^m/\tau\right) = -\log \frac{\exp\left((\hat{\boldsymbol{z}}_j^{l,(i)})^m \cdot (\hat{\boldsymbol{p}}_+^l)^m/\tau\right)}{\sum_{c=1}^{C}\exp\left((\hat{\boldsymbol{z}}_j^{l,(i)})^m \cdot (\hat{\boldsymbol{p}}_c^l)^m/\tau\right)}$. The proposed $\mathcal{L}_{\text{Hi-MPC}^h}$ in Eq. (5.9) inherits from $\mathcal{L}_{\text{Hi-MPC}}$ (see Eq. (5.5)) and combines the proposed HSM mechanism to adaptively infer the informative importance $\overline{\delta}(\cdot)$ of F skeletons in each sequence to enhance the proposed hierarchical skeleton meta-prototype contrastive learning. Instead of directly leveraging sequence-level skeleton representations $\boldsymbol{v}^{l,(i)}$ for Hi-MPC (see Sec. 5.2.2), the proposed $\mathcal{L}_{\text{Hi-MPC}^h}$ can take advantage of finer-grained pattern information contained in *each key skeleton* and corresponding hierarchical representations to mine more discriminative skeleton features for person re-ID tasks.

5.2.4 Multi-Level Skeleton Meta-Representation

To combine different-level body and motion semantics embedded in hierarchical skeleton representations and aggregate typical skeleton features learned from different meta-transformed subspaces, we propose to construct ***multi-level skeleton meta-representation (MSMR)*** as the final skeleton representation with:

$$\boldsymbol{V} = [\boldsymbol{V}^1; \boldsymbol{V}^2; \boldsymbol{V}^3] = [\frac{1}{M}\sum_{m=1}^{M}(\hat{\boldsymbol{v}}^1)^m; \frac{1}{M}\sum_{m=1}^{M}(\hat{\boldsymbol{v}}^2)^m; \frac{1}{M}\sum_{m=1}^{M}(\hat{\boldsymbol{v}}^3)^m]. \tag{5.10}$$

In Eq. (5.10), $\boldsymbol{V}^l \in \mathbb{R}^h$ denotes the l^{th} level skeleton meta-representation that aggregates features of skeleton meta-instances $(\hat{\boldsymbol{v}}^l)^m$ learned from M meta-transformed subspaces, $[;;]$ represents feature concatenation, and $\boldsymbol{V} \in \mathbb{R}^{3h}$ is the proposed MSMR that combines different-level skeleton meta-representations for person re-ID.

5.2.5 Workflow Overview of Hi-MPC

The computation flow of our approach can be summarized as: $\boldsymbol{S} \rightarrow \hat{\boldsymbol{S}}$ (Sec. 5.2.1) $\rightarrow \boldsymbol{v}$ (Sec. 5.2.2) $\rightarrow \hat{\boldsymbol{v}} \rightarrow \hat{\boldsymbol{p}} \rightarrow \hat{\text{y}}^m$ (Sec. 5.2.3) $\rightarrow \overline{\delta}((\hat{\boldsymbol{z}}_j)^m) \rightarrow \mathcal{L}_{\text{Hi-MPC}^h}$. As illustrated in Algorithm 1, we perform hierarchical skeleton meta-prototype contrastive learning by minimizing $\mathcal{L}_{\text{Hi-MPC}^h}$, so as to optimize the encoder $\psi(\cdot)$ and meta-transformation heads to learn effective skeleton representations in an unsupervised manner. During the optimization process, clustering and contrastive learning are

alternated to encourage better skeleton representation learning with more reliable clusters. For the person re-ID task, we encode the probe set $\Phi^{\mathcal{P}}$ into MSMR, $\{\boldsymbol{V}_i^{\mathcal{P}}\}_{i=1}^{N_2}$, and match it with corresponding representations, $\{\boldsymbol{V}_j^{\mathcal{G}}\}_{j=1}^{N_3}$, of the same identity in the gallery set $\Phi^{\mathcal{G}}$ using Euclidean distance.

5.3 Experiments

5.3.1 Experimental Setups

5.3.1.1 Datasets

We validated the effectiveness of our approach on four skeleton-based person re-ID benchmarks: *Kinect Gait Biometry Dataset (KGBD)* [24], *BIWI RGBD-ID Dataset (BIWI)* [1], *KS20 VisLab Multi-View Kinect Skeleton Dataset (KS20)* [119], *IAS-Lab RGBD-ID Dataset (IAS)* [120], and a large-scale RGB video based multi-view gait dataset: *CASIA-B* [121]. The original CASIA-B dataset does not contain 3D skeleton data, and we follow [25] to exploit pre-trained pose estimation models to extract 3D skeletons from RGB videos of CASIA-B, so as to evaluate the performance of our approach on RGB-estimated skeletons. As detailed in Table 4.1, these five datasets contain skeleton data of 164, 50, 20, 11, and 124 different identities, respectively. The full description and visual samples of datasets are provided in Sec. 4.3.1.1.

5.3.1.2 Probe and Gallery Settings

We follow the commonly-used settings of probe and gallery in the literature [5]: For the BIWI and IAS datasets, as different testing sets are non-overlapped and contain all pedestrians under different scenes, we evaluate our approach on each testing set by setting it as the probe while the other one is adopted as the gallery. The KGBD dataset contains different skeleton videos (*i.e.*, long skeleton sequences) of each pedestrian with varying numbers of walking rounds. Since no training/testing splits are given, we randomly choose one skeleton video of each person to split skeleton sequences and construct the probe set, and equally divide the remaining videos to build the training set and gallery set. The KS20 dataset collects skeleton

data of pedestrians from five different viewpoints, including 0°, 30°, 90°, 130°, and 180°. We employ different splitting setups to evaluate the multi-view person re-ID performance of our approach. For Random View Evaluation (RVE), one sequence is randomly selected from each viewpoint as the probe sequence and the remaining skeleton sequences are equally divided into gallery and training sequences. In Cross-View Evaluation (CVE) setup, we match persons across views, *i.e.*, matching between two different views that are not involved in training. We set sequences from two different viewpoints as the probe set and gallery set, respectively, while adopting sequences from the remaining viewpoints as the training set. The CASIA-B dataset contains sequences of 124 individuals under 11 different views and 3 conditions—pedestrians wearing a bag ("Bags"), wearing a coat ("Clothes"), and without any coat or bag ("Normal"). We follow the person re-ID protocols in [5, 13] (detailed in Sec. 4.3.1.3) to evaluate the proposed skeleton-based approach on CASIA-B. Experiments with each setup are repeated for multiple times and the average performance is reported in this work.

5.3.1.3 Implementation Details

Dataset Preprocessing Setups. To avoid ineffective skeleton recording, we discard the first and last 10 skeleton frames of each original skeleton sequence. For KS20, KGBD, BIWI, and IAS datasets, all skeleton sequences are normalized by subtracting the spine joint position from each joint of the same skeleton so that the skeleton is translation invariant [124]. Then, we spilt all normalized skeleton sequences in the training sets into multiple shorter skeleton sequences (*i.e.*, $\boldsymbol{S}_{1:f}$) with length f by a step of $\frac{f}{2}$, which aims to obtain as many 3D skeleton sequences as possible to train our approach. We split all skeleton sequences in the gallery and probe sets into shorter and non-overlapping sequences with length f. Unless explicitly specified, the skeleton sequence $\boldsymbol{S}_{1:f}$ in our work refers to those split and normalized sequences used in learning, rather than those original skeleton sequences provided by datasets. We follow the data augmentation strategy used in [4, 7, 8] to sample more sequences for different identities in the training set, and train our approach with randomly shuffled and unlabeled skeleton sequences.

Model Parameter Setups. As shown in Fig. 5.4, Fig. 5.5 and Fig. 5.6, the numbers of nodes in the three level skeleton representations are J (joint-level), 10 (component-level), and 5 (limb-level). We concatenate 3D coordinates of all nodes

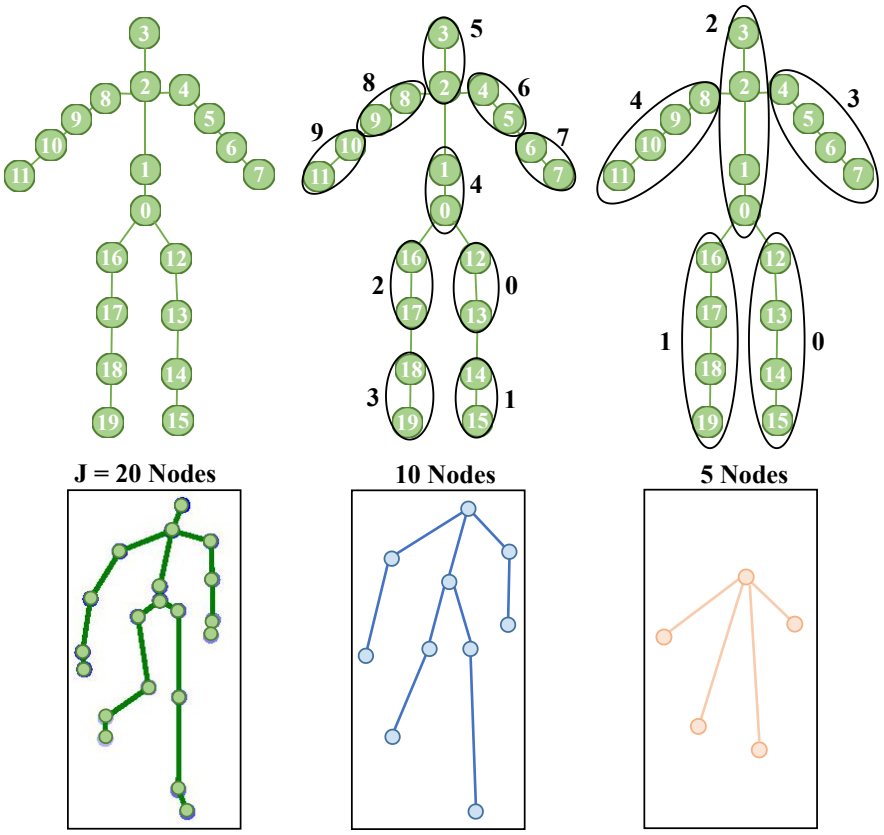


FIGURE 5.4: Node indices for joint-level (20 joints), component-level (10 nodes), and limb-level (5 nodes) representations of skeletons in IAS, BIWI and KGBD datasets. We spatially group body joints to be a more abstract body component in their central position (*i.e.*, the average position of all body joints in a group). The corresponding hierarchical skeleton representations are shown in the bottom row.

at each level. The skeleton sequence length F on four skeleton-based datasets (IAS, KS20, BIWI, KGBD) is set to 6 following [4] for a fair comparison with existing methods. As to CASIA-B, it is a large-scale dataset with roughly estimated skeleton data from RGB frames, which is intrinsically different from the previous datasets. We adopt a longer sequence length $F = 40$. The embedding size for skeleton representations is $h = h_1 = h_2 = 256$ for all datasets. We follow [5] to employ the DBSCAN algorithm [6] for clustering, as it can group feature-similar skeleton instances and discover semantic clusters with arbitrary shapes while not requiring pre-defined cluster number. We empirically set maximum distance $\epsilon = 0.6$ (KGBD, BIWI-S), $\epsilon = 0.8$ (KS20, IAS, BIWI-W), $\epsilon = 0.75$ (CASIA-B), and adopt minimum amount of samples $a_{min} = 4$ for KGBD and $a_{min} = 2$ for the other datasets. We follow [29, 125] to construct the commonly used Jaccard distance matrix to perform clustering, and discard all outliers in different clustered instance sets, *i.e.*,

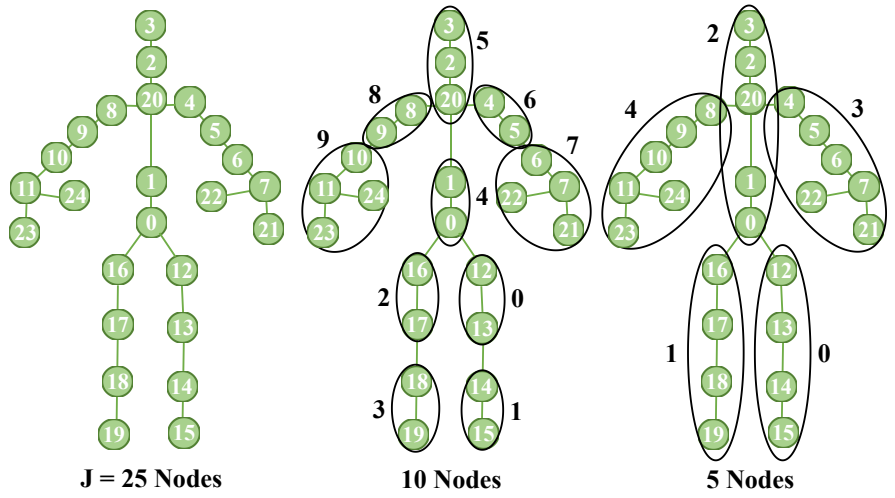


FIGURE 5.5: Node indices for joint-level (25 joints), component-level (10 nodes), and limb-level (5 nodes) representations of skeletons in KS20 dataset.

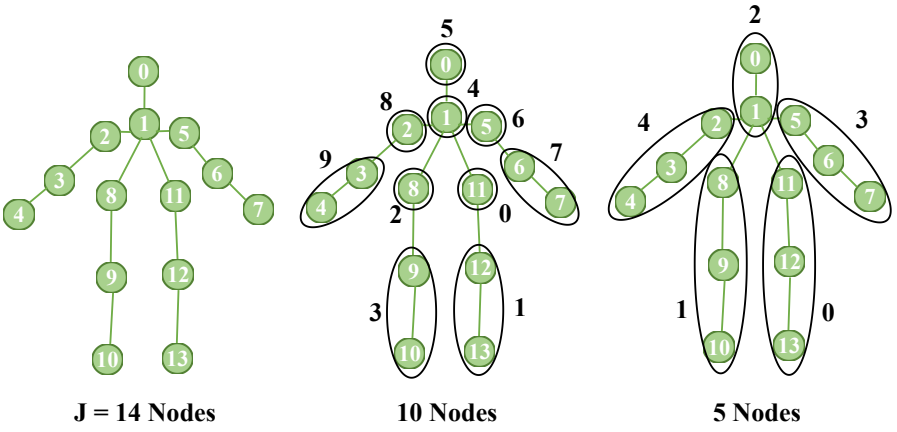


FIGURE 5.6: Node indices for joint-level (14 joints), component-level (10 nodes), and limb-level (5 nodes) representations of skeletons in CASIA-B dataset.

discard the union of all outliers, to perform contrastive learning. We employ the Adam optimizer with learning rate 0.00035 for all datasets. The batch size is set to 256 for all datasets. To avoid over-fitting and achieve better generalization performance, we adopt Early Stopping [126] with a patience of 50 epochs (*i.e.*, stop the training of model after no improvement in 50 continuous epochs). Interested readers can access our source codes at https://github.com/Kali-Hac/Hi-MPC to get more details.

Method Comparison Setups. For all methods compared in our experiments, we select optimal model parameters for training, and use their pre-defined skeleton descriptors or pre-trained skeleton representations for person re-ID. It is worth noting that our re-implementations of some existing models get performance with

slight variations, and the results are basically the same or even better than that presented in the original papers under different random model initializations. For a fair comparison, we follow [4, 5] to report the average performance of all methods. For the supervised fine-tuning of existing models, we attach an MLP network with one hidden layer to the end of original models, and train the MLP network on the frozen pre-trained representations with the supervision of labels. Then the feature representations before the last fully-connected layer are extracted for person re-ID. Our approach is trained with only unlabeled skeleton data without using any post-processing technique, *e.g.*, re-ranking [127] or multi-query fusion [49]. To perform person re-ID, we exploit the approach to encode hierarchical representations of each original skeleton sequence of the probe set $\Phi^{\mathcal{P}}$ into corresponding multi-level skeleton meta-representation (MSMR), $\{\boldsymbol{V}_i^{\mathcal{P}}\}_{i=1}^{N_2}$, and match it with representations, $\{\boldsymbol{V}_j^{\mathcal{G}}\}_{j=1}^{N_3}$, of the same identity in the gallery set $\Phi^{\mathcal{G}}$ using Euclidean distance. In the ablation study, we use the concatenation of raw skeleton sequences (*i.e.*, normalized 3D coordinates of body joints) as the baseline.

5.3.1.4 Evaluation Metrics

We compute the Cumulative Matching Characteristics (CMC) curve and adopt Rank-1 accuracy (R_1), Rank-5 accuracy (R_5), and Rank-10 accuracy (R_{10}) as performance metrics [5]. R_1, R_5, and R_{10} are computed as the ratios of probe sequences matching the gallery sequences with correct identities when the candidate gallery sequences are the top 1, top 1 to 5, and top 1 to 10 most similar sequences to the probe sequence. Mean Average Precision (mAP) [49] is also used to quantitatively evaluate the overall performance of our approach.

5.3.2 Empirical Evaluation

We compare our approach with state-of-the-art self-supervised and unsupervised skeleton-based person re-ID methods on BIWI, IAS, KGBD, and KS20 datasets, as shown in Tables 5.1 and 5.2. The latest supervised skeleton-based person re-ID methods [7, 25] and representative hand-crafted person re-ID methods [1, 2] are also included to provide a comprehensive comparison.

TABLE 5.1: Performance comparison with state-of-the-art skeleton-based methods on BIWI-S, BIWI-W, and IAS-A. ‡ indicates employing supervised fine-tuning and **Bold** denotes the best cases among self-supervised/unsupervised methods.

		BIWI-S				BIWI-W				IAS-A			
Types	**Methods**	$\mathbf{R_1}$	$\mathbf{R_5}$	$\mathbf{R_{10}}$	**mAP**	$\mathbf{R_1}$	$\mathbf{R_5}$	$\mathbf{R_{10}}$	**mAP**	$\mathbf{R_1}$	$\mathbf{R_5}$	$\mathbf{R_{10}}$	**mAP**
Hand-crafted	D_{13} [1]	28.3	53.1	65.9	13.1	14.2	20.6	23.7	17.2	40.0	58.7	67.6	24.5
	D_{16} [2]	32.6	55.7	68.3	16.7	17.0	25.3	29.6	18.8	42.7	62.9	70.7	25.2
Supervised	PoseGait [25]	14.0	40.7	56.7	9.9	8.8	23.0	31.2	11.1	28.4	55.7	69.2	17.5
	‡SGELA [4]	29.2	65.2	73.8	23.5	13.9	15.3	16.7	22.9	18.0	32.1	46.2	13.5
	MG-SCR [7]	20.1	46.9	64.1	7.6	10.8	20.3	29.4	11.9	36.4	59.6	69.5	14.1
	‡SM-SGE [8]	34.8	60.6	71.5	12.8	16.7	31.0	40.2	18.7	38.5	63.2	73.9	15.0
Self-supervised /Unsupervised	AGE [3]	25.1	43.1	61.6	8.9	11.7	21.4	27.3	12.6	31.1	54.8	67.4	13.4
	SGELA [4]	25.8	51.8	64.4	15.1	11.7	14.0	14.7	19.0	16.7	30.2	44.0	13.2
	SM-SGE [8]	31.3	56.3	69.1	10.1	13.2	25.8	33.5	15.2	34.0	60.5	71.6	13.6
	SimMC [5]	41.7	66.6	76.8	12.3	24.5	36.7	44.5	19.9	44.8	65.3	72.9	18.7
	Hi-MPC^h (Ours)	**47.5**	**70.3**	**78.6**	**17.4**	**27.3**	**40.3**	**48.8**	**22.6**	**45.6**	**67.3**	**75.4**	**23.2**

TABLE 5.2: Performance comparison with state-of-the-art skeleton-based methods on IAS-B, KGBD,, and KS20. ‡ indicates employing supervised fine-tuning and **Bold** denotes the best cases among self-supervised/unsupervised methods.

		IAS-B				KGBD				KS20			
Types	**Methods**	$\mathbf{R_1}$	$\mathbf{R_5}$	$\mathbf{R_{10}}$	**mAP**	$\mathbf{R_1}$	$\mathbf{R_5}$	$\mathbf{R_{10}}$	**mAP**	$\mathbf{R_1}$	$\mathbf{R_5}$	$\mathbf{R_{10}}$	**mAP**
Hand-crafted	D_{13} [1]	43.7	68.6	76.7	23.7	17.0	34.4	44.2	1.9	39.4	71.7	81.7	18.9
	D_{16} [2]	44.5	69.1	80.2	24.5	31.2	50.9	59.8	4.0	51.7	77.1	86.9	24.0
Supervised	PoseGait [25]	28.9	51.6	62.9	20.8	50.6	67.0	72.6	13.9	49.4	80.9	90.2	23.5
	‡SGELA [4]	23.6	42.9	51.9	14.8	43.7	58.7	65.0	7.1	49.7	67.0	77.1	22.2
	MG-SCR [7]	32.4	56.5	69.4	12.9	44.0	58.7	64.6	6.9	46.3	75.4	84.0	10.4
	‡SM-SGE [8]	44.3	68.2	77.5	14.9	43.2	58.6	64.6	7.5	49.8	78.1	85.2	11.7
Self-supervised /Unsupervised	AGE [3]	31.1	52.3	64.2	12.8	2.9	5.6	7.5	0.9	43.2	70.1	80.0	8.9
	SGELA [4]	22.2	40.8	50.2	14.0	38.1	53.5	60.0	4.5	45.0	65.0	75.1	21.2
	SM-SGE [8]	38.9	64.1	75.8	13.3	38.2	54.2	60.7	4.4	45.9	71.9	81.2	9.5
	SimMC [5]	46.3	68.1	77.0	22.9	54.9	66.2	70.6	**11.7**	66.4	80.7	87.0	**22.3**
	Hi-MPC^h (Ours)	**48.2**	**70.2**	**77.8**	**25.3**	**56.9**	**70.2**	**75.1**	10.2	**69.6**	**83.5**	**87.1**	22.0

5.3.2.1 Comparison with Self-Supervised and Unsupervised State-of-the-Art Methods

The proposed approach shows significant performance improvements over existing self-supervised and unsupervised methods on different datasets. Compared with AGE [3] and SGELA [4] that learn skeleton features from only joint-level representations, our approach achieves higher person re-ID performance on all datasets with an improvement of 14.5-54.0% for Rank-1 accuracy, 12.5-64.6% for Rank-5 accuracy, 7.1-67.6% for Rank-10 accuracy, and 0.8-13.1% for mAP. These improvements indicate that the proposed approach can capture more discriminative features from multiple skeleton levels for person re-ID. By mining key skeletons and exploiting more informative patterns of different levels with the proposed HSM mechanism, our approach significantly outperforms the SM-SGE framework [8] that directly

TABLE 5.3: Cross-view person re-ID performance comparison with state-of-the-art self-supervised and unsupervised methods with CVE setup of KS20. $0°, 30°, 90°, 130°$, and $180°$ denote different views of probe or gallery sets.

	Gallery Views	0°				30°				90°				130°				180°			
Probe Views		**R_1**	**R_5**	**R_{10}**	**mAP**	**R_1**	**R_5**	**R_{10}**	**mAP**	**R_1**	**R_5**	**R_{10}**	**mAP**	**R_1**	**R_5**	**R_{10}**	**mAP**	**R_1**	**R_5**	**R_{10}**	**mAP**
0°	AGE [3]	46.7	74.2	83.5	22.5	11.0	35.7	47.5	10.0	8.1	29.9	47.5	9.2	7.5	26.7	43.5	8.4	7.0	23.0	37.4	8.2
	SGELA [4]	76.2	89.6	92.8	37.1	15.1	27.3	35.1	19.9	10.1	27.5	40.9	18.2	10.7	21.5	29.3	18.0	15.4	25.8	38.0	12.6
	SM-SGE [8]	58.4	84.7	92.2	27.7	17.2	50.0	63.3	10.8	7.2	21.9	39.1	10.5	4.4	19.4	34.7	9.3	10.0	23.8	33.1	9.4
	SimMC [5]	84.4	97.3	99.2	61.2	**37.9**	59.4	67.6	24.8	28.9	50.8	62.9	**27.1**	23.3	**43.0**	52.9	**20.3**	15.2	29.3	**45.7**	**14.5**
	Hi-MPC^h (Ours)	**92.2**	**98.4**	**99.6**	**66.9**	35.6	**61.7**	**72.7**	**27.3**	**36.7**	**58.6**	**73.1**	25.9	**24.6**	41.8	**53.5**	16.4	**17.2**	**30.1**	43.8	13.2
30°	AGE	10.1	42.8	57.8	8.8	52.3	82.7	91.5	25.0	15.0	35.6	58.5	8.8	10.1	24.2	41.8	8.1	7.8	24.2	34.3	8.3
	SGELA	13.1	19.6	22.6	19.4	70.9	88.2	91.8	40.5	11.8	24.5	36.3	16.5	6.9	22.6	31.7	15.4	9.2	15.4	22.9	13.9
	SM-SGE	18.1	48.4	65.0	11.5	60.2	82.0	89.8	28.2	12.5	27.2	35.3	10.7	7.5	23.4	33.8	10.6	8.8	27.2	39.1	10.5
	SimMC	30.8	66.2	74.6	20.7	91.8	97.4	98.2	**67.8**	36.8	55.1	67.6	**29.9**	16.4	30.5	40.2	**20.4**	16.2	36.7	**56.6**	12.7
	Hi-MPC^h (Ours)	**33.2**	**66.4**	**75.8**	**24.4**	**93.8**	**97.7**	**98.8**	66.5	**37.5**	**62.1**	**71.5**	25.1	**19.5**	**34.8**	**50.0**	17.8	**16.8**	**37.5**	47.7	**14.3**
90°	AGE	7.5	27.3	43.2	8.7	9.0	28.5	44.1	9.3	57.4	81.4	90.7	19.2	13.8	41.1	57.1	9.0	7.8	30.0	46.0	8.3
	SGELA	9.6	19.8	29.7	16.4	10.8	15.6	20.4	17.5	48.4	75.7	86.5	31.6	17.1	35.7	43.0	22.0	13.5	23.4	31.8	21.3
	SM-SGE	19.1	33.1	48.1	12.4	23.1	40.6	57.4	11.5	72.2	89.1	92.8	24.9	20.9	48.4	69.4	12.8	19.4	36.9	51.6	11.3
	SimMC	**26.2**	44.9	50.8	11.9	41.4	64.1	75.4	27.3	96.7	**100**	**100**	73.1	60.9	**81.6**	**88.7**	**45.0**	25.8	48.4	64.5	15.4
	Hi-MPC^h (Ours)	**26.2**	**47.7**	**62.1**	**23.0**	**50.8**	**71.5**	**83.2**	**34.0**	**97.3**	**100**	**100**	**73.9**	**61.7**	80.5	84.8	42.2	**33.2**	**65.2**	**80.1**	**23.1**
130°	AGE	6.7	21.3	34.7	8.2	7.9	23.4	38.9	8.9	15.2	35.9	54.4	9.2	45.3	70.5	82.1	18.7	11.3	37.1	50.2	8.9
	SGELA	5.8	18.8	28.0	14.2	11.6	15.5	20.7	16.8	17.6	47.1	53.2	24.5	59.6	81.5	89.1	36.8	17.0	29.8	32.5	23.0
	SM-SGE	8.4	24.4	37.8	10.4	12.9	26.6	36.3	10.9	24.1	53.4	66.3	12.9	64.4	85.9	95.0	25.5	17.8	40.9	59.1	12.1
	SimMC	18.0	32.4	48.8	14.2	**24.2**	44.9	59.4	15.7	60.2	78.1	86.7	45.2	92.5	98.8	**99.2**	71.5	30.1	55.1	66.8	18.8
	Hi-MPC^h (Ours)	**19.9**	**39.1**	**55.1**	**20.3**	20.7	**50.8**	**68.4**	**21.9**	**62.1**	**80.5**	**87.5**	**45.8**	**93.4**	**99.2**	**99.2**	**72.8**	**36.3**	**61.3**	**75.8**	**26.3**
180°	AGE	7.9	17.7	32.6	8.1	5.2	22.4	33.4	8.3	10.5	25.6	34.0	8.2	11.6	33.1	52.9	8.8	47.1	72.4	82.6	22.6
	SGELA	14.0	29.1	39.2	21.3	11.9	20.6	25.9	**17.3**	18.6	37.8	49.7	19.4	22.7	45.9	55.2	20.7	74.5	92.7	95.1	38.3
	SM-SGE	5.6	20.0	30.6	8.5	6.6	22.7	31.6	8.6	13.8	34.1	45.6	9.4	10.3	37.5	56.6	10.4	51.9	79.7	87.8	25.6
	SimMC	**19.1**	**39.5**	**48.8**	15.3	14.1	28.1	40.6	14.0	25.8	44.1	53.5	**29.8**	36.9	61.3	74.6	**32.0**	91.0	**97.7**	98.4	**59.5**
	Hi-MPC^h (Ours)	16.8	34.4	47.6	**15.6**	**14.8**	**30.9**	**44.5**	14.8	**30.9**	**55.5**	**68.0**	25.8	**37.1**	**62.5**	**79.3**	29.5	**93.0**	97.3	**98.8**	58.0

combines four-scale skeletons by 9.3-23.7% for Rank-1 accuracy and 5.8-12.5% for mAP on different datasets. Note that we follow [4, 5] to report the average performance of all methods for a fair comparison while there exist slight performance variations in practice. In this context, the proposed approach is still superior to the latest skeleton contrastive learning framework SimMC [5] in most cases, as shown in Tables 5.1 and 5.2. Although SimMC obtains slightly higher mAP on KS20 and KGBD, our approach can achieve better overall performance in terms of Rank-1, Rank-5, Rank-10, and mAP on the datasets that contain frequent changes of appearances or scenes (BIWI-S, BIWI-W, IAS-A, IAS-B). Considering that these changes could induce more random perturbations or noises in the collection of skeletons, the results suggest that our approach is more robust than SimMC for learning effective skeleton representations under different conditions. Moreover, our approach is more efficient than most skeleton-based person re-ID methods in terms of the model size and computational complexity (see Sec. 5.4.6).

Our approach is also evaluated on the cross-view person re-ID scenarios (*i.e.*, CVE setup) of KS20. As presented in Table 5.3, the proposed Hi-MPC^h outperforms state-of-the-art self-supervised and unsupervised counterparts by an average margin of 6.4 to 44.4% for Rank-1 accuracy and 3.0 to 48.7% for mAP on all views,

TABLE 5.4: Ablation study with different configurations: Direct prototype contrastive learning (DPC), meta-prototype contrastive learning (MPC), and hard skeleton mining mechanism (HSM). "Hi" denotes adopting hierarchical skeleton representations and exploiting the proposed MSMR for person re-ID. "+" indicates using the corresponding component.

ID	Configurations	BIWI-S		BIWI-W		IAS-A		IAS-B		KGBD		KS20	
		R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
1	Baseline	24.8	9.3	10.9	14.1	29.4	13.8	30.2	13.3	20.5	4.4	17.0	9.5
2	+ DPC	38.3	10.7	19.9	19.7	35.4	16.3	35.4	16.6	53.7	8.5	63.3	17.6
3	+ MPC	39.8	13.1	22.4	19.3	38.4	17.0	37.3	14.4	53.2	8.1	63.9	18.5
4	+ Hi + MPC	40.4	12.8	24.2	21.1	42.2	19.4	38.2	18.9	55.3	8.7	66.4	18.6
5	+ MPC + HSM	44.9	14.2	23.7	21.0	40.0	17.8	42.8	23.1	55.8	9.5	66.2	19.0
6	+ Hi + MPC + HSM	47.5	17.4	27.3	22.6	45.6	23.2	48.2	25.3	56.9	10.2	69.6	22.0

and also surpasses the latest SimMC framework [5] in most probe-gallery matching views. This suggests that our model is more effective on learning generalized (*i.e.*, view-independent) skeleton representations with higher robustness to viewpoint changes for cross-view person re-ID.

5.3.2.2 Comparison with Hand-Crafted and Supervised State-of-the-Art Methods

The results in Tables 5.1 and 5.2 show that our model achieves better person re-ID performance than the representative hand-crafted methods D_{13} [1] and D_{16} [2] that utilize numerous anthropometric skeleton descriptors by 3.7-39.9% for Rank-1 accuracy and 0.7-8.3% for mAP on four of the six testing sets (BIWI-S, BIWI-W, IAS-B, KGBD). Although they achieve competitive mAP on datasets with frequent appearance and viewpoint changes (IAS-A, KS20), our approach can obtain higher overall performance in terms of Rank-1 accuracy (2.9-30.2%), Rank-5 accuracy (4.4-11.8%), and Rank-10 accuracy (0.2-7.8%). Notably, the proposed unsupervised approach markedly outperforms supervised state-of-the-art models PoseGait [25] and MG-SCR [7] on almost all datasets. Interestingly, with extra labels to fine-tune SGELA [4] and SM-SGE [8], those methods still perform poorly on many datasets. In contrast, our approach achieves better and more stable performance on all datasets without using any skeletal annotation. This shows that Hi-MPC^h has good generality with greater potential for use in practical person re-ID scenarios.

5.4 Further Analysis

5.4.1 Ablation Study

We conduct ablation study to demonstrate the contribution of each component in our approach. The raw skeleton sequences, *i.e.*, concatenation of 3D coordinates of body joints, are adopted as the baseline for comparison. As reported in Table 5.4, employing direct skeleton prototype contrastive (DPC) learning (ID = 2) significantly improves the person re-ID performance of the baseline especially on more challenging datasets such as KGBD (up to 33.2% for Rank-1 accuracy) and KS20 (up to 46.3% for Rank-1 accuracy). This suggests that prototypes are indeed more representative skeleton features than raw 3D skeletons and DPC plays a crucial role in exploiting such discriminative features from different individuals. The proposed meta-prototype contrastive (MPC) learning (ID = 3) performs better than DPC (ID = 2) in almost all cases when applied to joint-level skeleton representations. Combining hierarchical skeleton representations for multi-level clustering and meta-prototype contrastive learning (ID = 4) achieves better performance than joint-level DPC (ID = 2) by up to 6.8% for Rank-1 accuracy and 3.1% for mAP on different datasets. Such results verify the effectiveness of the proposed Hi-MPC, as it can capture richer body and motion patterns via exploiting the most typical skeleton features from various levels. The effects of different level skeleton representations will be also discussed in Sec. 5.4.3. The proposed approach (Hi-MPC^h) employing the hard skeleton mining (HSM) mechanism (ID = 6) achieves consistent performance improvement of 1.6-10.0% for Rank-1 accuracy and 1.2-6.4% for mAP compared with naïve Hi-MPC (ID = 4) on all datasets. Adding HSM (ID = 5) also improves the performance of MPC (ID = 3), which suggests the general validity of HSM for both single-level and hierarchical skeleton contrastive learning. These results further suggest that HSM can facilitate mining key skeletons to learn highly informative and valuable skeletal patterns during skeleton meta-prototype contrastive learning for person re-ID. Further visualization and analysis of the HSM mechanism are provided in Sec. 5.4.8.

TABLE 5.5: Performance comparison with appearance-based and skeleton-based methods on CASIA-B. "Clothes-Normal" represents the probe set under "Clothes" condition and gallery set under "Normal" condition. ♣ refers to appearance-based methods and ‡ represents requiring label information for training. "—" indicates no published result.

Probe-Gallery	**Normal-Normal**				**Bags-Bags**				**Clothes-Clothes**				**Clothes-Normal**				**Bags-Normal**			
Methods	R_1	R_5	R_{10}	**mAP**	R_1	R_5	R_{10}	**mAP**	R_1	R_5	R_{10}	**mAP**	R_1	R_5	R_{10}	**mAP**	R_1	R_5	R_{10}	**mAP**
‡LMNN♣ [9]	3.9	22.7	36.1	—	18.3	38.6	49.2	—	17.4	35.7	45.8	—	11.6	12.6	17.8	—	23.1	37.1	44.4	—
‡ITML♣ [10]	7.5	22.2	34.2	—	19.5	26.0	33.7	—	20.1	34.4	43.3	—	10.3	24.5	36.1	—	21.8	30.4	36.3	—
‡ELF♣ [11]	12.3	35.6	50.3	—	5.8	25.5	37.6	—	19.9	43.9	56.7	—	5.6	16.0	26.3	—	17.1	30.0	37.9	—
SDALF♣ [12]	4.9	27.0	41.6	—	10.2	33.5	47.2	—	16.7	42.0	56.7	—	11.6	19.4	27.6	—	22.9	30.1	36.1	—
‡Score-based MLR♣ [13]	13.6	48.7	63.7	—	13.6	48.7	63.7	—	13.5	48.6	63.9	—	9.7	27.8	45.1	—	14.7	32.6	50.2	—
‡Feature-based MLR♣ [13]	16.3	43.4	60.8	—	18.9	44.8	59.4	—	25.4	53.3	68.9	—	20.3	42.6	**56.9**	—	31.8	53.6	64.1	—
AGE [3]	20.8	29.3	34.2	3.5	37.1	56.2	67.0	9.8	35.5	54.3	65.3	9.6	14.6	33.0	42.7	3.0	32.4	51.2	60.1	3.9
SM-SGE [8]	50.2	73.5	81.9	6.6	26.6	49.0	59.4	9.3	27.2	51.4	63.2	9.7	10.6	26.3	35.9	3.0	16.6	36.8	47.5	3.5
SGELA [4]	71.8	87.5	91.4	9.8	48.1	69.5	77.7	16.5	51.2	73.8	81.5	7.1	15.9	30.8	40.6	4.7	36.4	57.1	64.6	6.7
SimMC [5]	80.8	92.3	93.7	10.8	69.1	86.6	91.3	16.5	68.0	88.1	**93.0**	**15.7**	25.6	43.8	54.0	**5.4**	42.0	59.8	68.9	7.1
Hi-MPC^h (Ours)	**85.5**	**94.9**	**95.8**	**11.2**	**71.2**	**87.5**	**92.1**	**17.0**	**70.2**	**88.5**	92.6	14.1	**27.2**	**45.0**	54.9	4.9	**50.1**	**65.5**	**72.1**	**7.5**

5.4.2 Evaluation on Model-Estimated Skeletons

In this section, we verify the generality of our skeleton-based approach under the large-scale RGB scenarios (CASIA-B). We leverage pre-trained pose estimation models [122, 123] to extract 3D skeleton data from RGB videos of CASIA-B, and evaluate the performance of our approach with the estimated skeleton data. As shown in Table 5.5, our approach achieves superior performance to most state-of-the-art skeleton-based methods (AGE [3], SM-SGE [8], SGELA [4]) with a significant margin of 11.3 to 64.7% for Rank-1 accuracy and 0.2 to 7.7% for mAP in different evaluation conditions of CASIA-B. Compared with the latest contrastive learning framework SimMC [5], our model performs better in most conditions, which justifies its effectiveness on learning more discriminative skeleton representations when applied to model-estimated skeleton data. Our skeleton-based approach also outperforms different representative classical appearance-based methods that employ visual metric learning with RGB-based appearances and textures (LMNN [9], ITML [10], ELF [11], SDALF [12]) or leverage both gait energy images and appearance features (MLR [13]) on different conditions. The highly competitive performance of the proposed approach on RGB-estimated skeletons demonstrates its generality and value for person re-ID under large-scale RGB-based scenarios.

TABLE 5.6: Performance of our approach with different level representations (joint-level, component-level, limb-level) and their combination (multi-level) on different datasets. The multi-level graph method SM-SGE is compared under the same setting of skeleton levels. Note: The joint-level, component-level (10 components), and limb-level (5 limbs) skeleton representations correspond to joint-scale, part-scale (10 parts), and body-scale (5 limbs) skeleton graphs in SM-SGE.

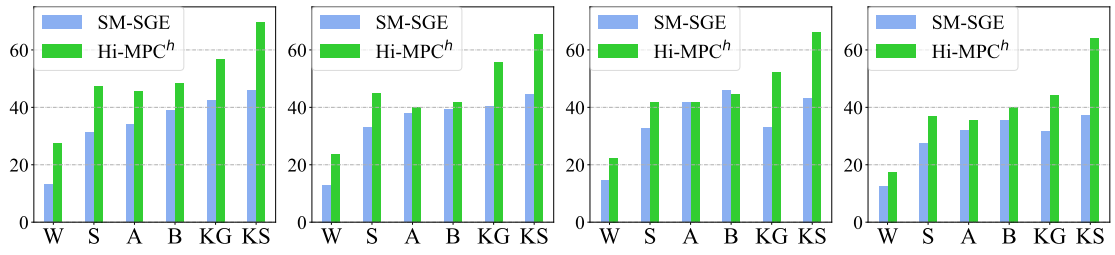
Levels	Methods	BIWI-S		BIWI-W		IAS-A		IAS-B		KS20		KGBD	
		R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
Multi-Level	SM-SGE	31.3	10.1	13.2	15.2	34.0	13.6	38.9	13.3	45.9	9.5	42.3	4.4
	Hi-MPC^h	47.5	17.4	27.3	22.6	45.6	23.2	48.2	25.3	69.6	22.0	56.9	10.2
Joint-Level	SM-SGE	33.0	10.0	12.9	14.9	37.8	14.7	39.4	13.9	44.7	10.2	40.2	4.3
	Hi-MPC^h	44.9	14.2	23.7	21.0	40.0	17.8	41.7	20.1	65.4	18.9	55.8	9.5
Component-Level	SM-SGE	32.8	11.1	14.5	16.5	41.6	15.2	45.9	15.8	43.2	9.8	33.0	4.1
	Hi-MPC^h	41.8	15.1	22.3	18.9	41.7	18.1	44.6	21.5	66.0	18.8	52.3	7.9
Limb-Level	SM-SGE	27.5	10.0	12.6	13.8	31.9	13.8	35.5	13.5	37.3	9.3	31.5	4.4
	Hi-MPC^h	36.9	12.4	17.5	16.3	35.5	17.7	40.1	19.3	64.1	17.9	44.3	5.7

TABLE 5.7: Performance of our approach on different datasets when solely exploiting limb-level (L), component-level (C), joint-level skeleton representations (J) or MSMR that combines all level skeleton representations for person re-ID. We compare our approach and the naïve Hi-MPC without using HSM. "✓" indicates using the corresponding configurations.

ID	L	C	J	HSM	BIWI-S		BIWI-W		IAS-A		IAS-B		KGBD		KS20	
					R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
1	✓				35.7	12.1	18.0	16.8	35.2	16.1	38.0	18.3	42.5	5.5	59.2	15.6
2		✓			38.9	12.2	21.8	18.1	36.9	16.0	38.0	19.8	50.0	7.0	59.8	16.4
3			✓		39.3	12.3	22.4	19.3	38.4	17.0	37.3	14.4	51.5	7.7	61.8	16.5
4	✓	✓	✓		40.4	12.8	24.2	21.1	42.2	19.4	38.2	18.9	55.3	8.7	65.4	18.2
5	✓			✓	36.9	12.4	17.5	16.3	35.5	17.7	40.1	19.3	44.3	5.7	64.1	17.9
6		✓		✓	41.8	15.1	22.3	18.9	41.7	18.1	44.6	20.5	52.3	7.9	66.0	18.8
7			✓	✓	44.9	14.2	23.7	21.0	40.0	17.8	41.7	20.1	55.8	9.5	65.4	18.9
8	✓	✓	✓	✓	47.5	17.4	27.3	22.6	45.6	23.2	48.2	25.3	56.9	10.2	69.6	22.0

5.4.3 Evaluation on Different Level Skeleton Representations

We evaluate the performance of limb-level, component-level, joint-level skeleton representations and their combination in the proposed approach. As shown in Table 5.7, different level skeleton representations (ID = 5-7) learned by our final approach can *individually* achieve highly competitive performance on different datasets, which suggests the great potential of higher level skeleton representations such as key limbs to be directly applied for person re-ID. Notably, the model exploiting joint-level (ID = 3, 7) and component-level skeleton representations (ID = 2, 6) achieves higher performance than using limb-level (ID = 1, 5) in most cases. This demonstrates the greater contribution of low-level skeleton representations, as they usually contain more specific positional and structural information of body



(A) M-Level (B) J-Level (C) C-Level (D) L-Level

FIGURE 5.7: Performance (Rank-1 accuracy) of our approach with different-level representations (Joint(J)-level, Component(C)-level, Limb(L)-level) and their combination (Multi(M)-level) on BIWI-W (W), BIWI-S (S), IAS-A (A), IAS-B (B), KGBD (KG), and KS20 (KS) testing sets. The latest multi-level method SM-SGE is compared under the same setting of skeleton levels.

parts than high-level ones to benefit learning more discriminative features for person re-ID. Furthermore, combining hierarchical skeleton representations of all levels (ID = 4, 8) attains the best person re-ID performance on different datasets when compared to single-level representations (ID = 1-3, 5-7), regardless of using HSM. Such results further verify the necessity of the proposed hierarchical skeleton representations, as the multi-level skeletal modeling can encourage mining more unique body and motion features for person re-ID, which is consistent with the analysis in [7, 8].

We also compare the performance of different-level skeleton representations learned by our approach with the latest multi-level graph framework SM-SGE [8]. As presented in Fig. 5.7, our approach not only significantly outperforms SM-SGE using multi-level representations, but also gains superior performance when applying the learned higher level representations for person re-ID on five of six testing sets. This further demonstrates the effectiveness of our approach on learning more useful skeleton features at various levels. It is interesting to observe that the component-level representations with a simpler body structure can perform comparably or even better than joint-level representations on half of datasets. This implies that the original skeletons of those datasets might contain redundant positional or structural information, which can be compressed and characterized with more concise and abstract skeleton representations to better achieve person re-ID.

TABLE 5.8: Generalized person re-ID performance of our approach with direct domain generalization (DG) from source datasets ("Source") to target datasets ("Target"). "UF" represents fine-tuning the source model with the unlabeled data of target datasets. **Bold** indicates that the model using "DG" or "UF" obtains better performance than the original one trained on the same dataset.

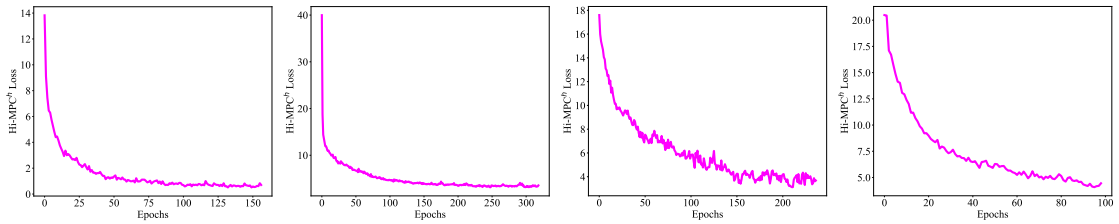
	Target	BIWI-S		BIWI-W		IAS-A		IAS-B		KGBD	
Source	Types	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
BIWI	DG	—	—	—	—	29.7	13.1	31.8	12.8	41.4	3.6
	UF	47.5	17.4	27.3	22.6	45.5	22.3	46.9	**28.0**	**57.3**	8.7
IAS	DG	26.0	8.9	9.8	11.4	—	—	—	—	41.0	3.3
	UF	45.9	**17.9**	**27.6**	**23.2**	45.6	23.2	48.2	25.3	46.2	4.3
KGBD	DG	31.5	10.2	13.4	15.2	27.6	12.1	27.0	12.4	—	—
	UF	**51.0**	17.3	22.8	20.0	**45.7**	20.1	**49.1**	**26.2**	56.9	10.2

5.4.4 Evaluation on Generalized Person Re-Identification

The pre-trained model of our approach can be transferred to different datasets with the same type of skeleton data (*i.e.*, containing the same number of joints) for the generalized person re-ID task. Here we take the BIWI, IAS, and KGBD datasets that contain the same type of skeletons for evaluation. Specifically, we exploit the model trained on the training set of a source dataset to perform person re-ID on the testing/probe set of target dataset (*i.e.*, direct domain generalization (DG)), and then further fine-tune the model with the unlabeled training data of target datasets (*i.e.*, unsupervised fine-tuning (UF)). As shown in Table 5.8, direct generalization is effective among different datasets, while unsupervised fine-tuning on the target dataset can further improve the person re-ID performance, which can be better than the model trained on the original dataset (see bold in Table 5.8). This demonstrates that our approach possesses generalization ability against domain shifts [135] and can be promisingly applied to other open person re-ID tasks. It is worth noting that training on different source datasets typically leads to different person re-ID performance on a new dataset, implying that an appropriate domain initialization or model pre-training of our model could be potentially exploited to facilitate better generalized person re-ID performance.

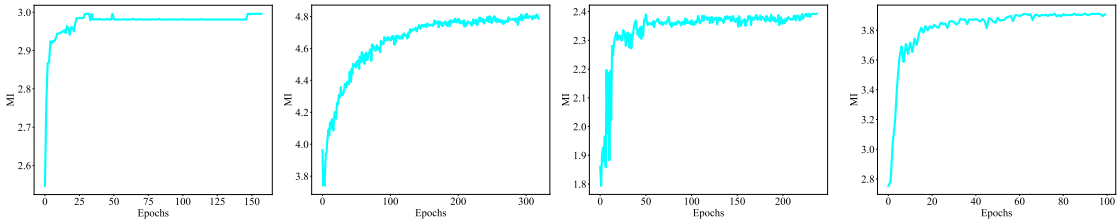
5.4.5 Feature Visualization

We provide a qualitative analysis with t-SNE [128] visualization of skeleton representations, which are compared with two skeleton-based state-of-the-art methods, SM-SGE [8] and SimMC [5]. As shown in Fig. 5.11c, the proposed MSMR



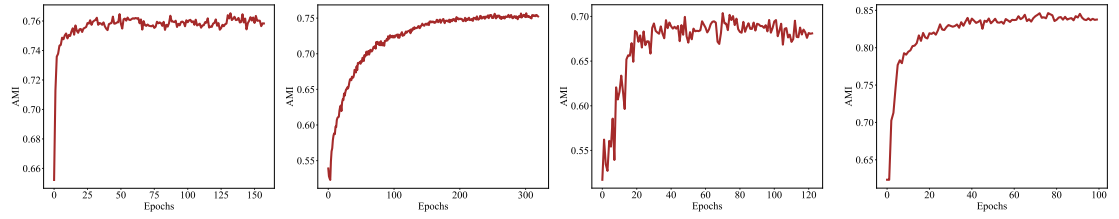
(A) KS20 (B) KGBD (C) IAS (D) BIWI

FIGURE 5.8: The model loss ($\mathcal{L}_{\text{Hi-MPC}^h}$) curves on the training sets of different datasets.



(A) KS20 (B) KGBD (C) IAS (D) BIWI

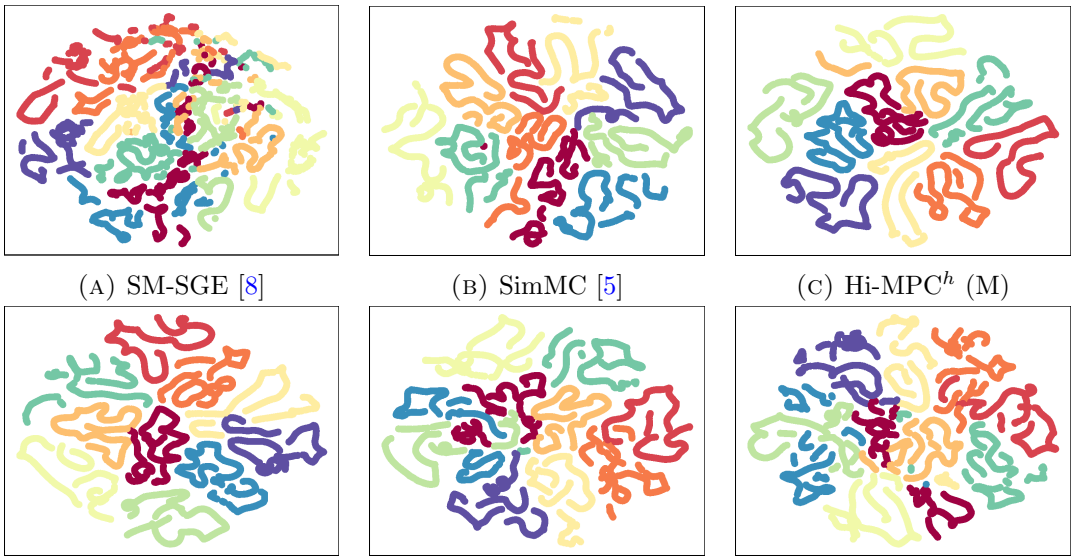
FIGURE 5.9: Mutual information (MI) between the clusters/pseudo classes generated by our approach and ground-trutth class labels on the training sets of different datasets.



(A) KS20 (B) KGBD (C) IAS (D) BIWI

FIGURE 5.10: Adjusted mutual information (AMI) score between the clusters/pseudo classes generated by our approach and ground-truth labels on the training sets of different datasets.

learned from our approach achieves higher inter-identity separation than representations learned from SM-SGE. Compared with SimMC, our method is able to simultaneously learn coarse-to-fine (*e.g.*, joint-level, component-level) skeleton representations with lower entropy, which forms evident identity groups in different levels, as shown in Fig. 5.11d and 5.11e. Such results not only suggest the effectiveness of Hi-MPC^h on learning discriminative representations (*i.e.*, differences between ground-truth classes) from *unlabeled* 3D skeleton data, but also demonstrate its ability on learning skeleton semantics (*e.g.*, identity-specific patterns) at different levels, which is consistent with the conclusions in Sec. 5.4.3. However, it can be that the limb-level skeleton representations with the much more abstract



(A) SM-SGE [8] (B) SimMC [5] (C) Hi-MPC^h (M)

(D) Hi-MPC^h (J) (E) Hi-MPC^h (C) (F) Hi-MPC^h (L)

FIGURE 5.11: t-SNE visualization of skeleton representations learned from SM-SGE (a), SimMC (b), and Hi-MPC^h ((c)-(f)) for the first 10 identities in BIWI. We visualize MSMR (M), joint-level (J), component-level (C), and limb-level representations (L) learned by Hi-MPC^h in (c), (d), (e), and (f), respectively. Features of different identities are shown in different colors.

TABLE 5.9: The number of network parameters (million (M)) and computational complexity (giga floating-point operations (GFLOPs)) of deep learning based methods. ‡ indicates employing supervised fine-tuning.

Types	**Methods**	**# Params**	**GFLOPs**
Supervised	PoseGait [25]	8.93M	121.60
	‡SGELA [4]	9.09M	7.48
	MG-SCR [7]	0.35M	6.60
	‡SM-SGE [8]	6.25M	23.92
Self-supervised /Unsupervised	AGE [3]	7.15M	37.37
	SGELA [4]	8.47M	7.47
	SM-SGE [8]	5.58M	22.61
	SimMC [5]	0.15M	0.99
	Hi-MPC^h (Ours)	3.32M	3.37

body structure are more difficult to be clustered in the t-SNE visualization (see Fig. 5.11f). This implies that too much information loss of skeleton positions, structures, and dynamics (*e.g.*, full dynamics of all joints) could reduce the model performance on learning recognizable pattern information of different identities.

TABLE 5.10: Performance of our approach with different meta-transformation heads (M = 1, 4, 8, 16).

M	BIWI-S		BIWI-W		IAS-A		IAS-B		KS20		KGBD	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
1	45.7	15.3	25.7	22.5	43.8	21.1	46.7	25.0	67.2	20.6	55.5	9.2
4	47.1	17.6	26.6	22.9	45.0	22.5	48.1	26.8	69.2	21.3	56.5	10.0
8	47.5	17.4	27.3	22.6	45.6	23.2	48.2	25.3	69.6	22.0	56.9	10.2
16	48.5	16.9	27.7	23.9	46.0	24.0	48.2	25.4	70.7	21.4	57.0	9.7

5.4.6 Model Efficiency

We report the model efficiency in terms of model size, *i.e.*, number of network parameters, and computational complexity for existing deep learning based methods. Both numbers of parameters and GFLOPs in the training of neural networks are counted by the Tensorflow platform [136]. For the model that possesses varying sizes and complexities on different datasets due to the changes of input data, we report the largest case. For models with direct supervised fine-tuning, we add the size of original models with the size of corresponding fine-tuning MLP network (Note that different output sizes of pre-trained skeleton representations influence the network size in the fine-tuning process). As shown in Table 5.9, the proposed approach possesses smaller model size and GFLOPs than most existing skeleton-based person re-ID methods (PoseGait [25], AGE [3], SGELA [4], SM-SGE [8]) while achieving state-of-the-art performance on all datasets as presented in our work.

The number of GFLOPs in Table 5.9 refers to computational complexity in the training of neural networks, which is the whole[2]/main computational complexity for deep learning methods. Like SimMC [5], our model needs extra matrix computation in the clustering process (*e.g.*, vector similarity query), where we employ the Faiss library [137] to optimize the overall complexity.

5.4.7 Analysis of Hyperparameters

5.4.7.1 Effects of Meta-Transformation Heads

Table 5.10 presents the effects of different numbers of meta-transformation heads.

[2]For representation learning methods without other learning processes (*e.g.*, clustering), the whole computational complexity of the model can be equivalent to the computational complexity of the used neural networks.

TABLE 5.11: Performance of our approach with different embedding sizes (h = 64, 128, 256, 512).

h	BIWI-S		BIWI-W		IAS-A		IAS-B		KS20		KGBD	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
64	45.1	16.3	25.1	21.6	44.0	21.7	46.5	24.5	69.2	21.4	54.7	8.7
128	47.0	17.5	27.5	22.9	45.0	22.9	49.3	26.9	69.7	20.8	55.1	9.5
256	47.5	17.4	27.3	22.6	45.6	23.2	48.2	25.3	69.6	22.0	56.9	10.2
512	47.0	15.7	26.2	22.3	42.5	21.8	49.5	27.3	70.7	22.0	55.8	9.2

TABLE 5.12: Performance of our approach when setting different minimum sample amount (a_{min} = 1, 2, 3, 4) for DBSCAN.

a_{min}	BIWI-S		BIWI-W		IAS-A		IAS-B		KS20		KGBD	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
1	46.6	17.6	26.6	21.0	44.7	22.1	48.5	26.1	68.4	21.1	54.7	9.6
2	47.5	17.4	27.3	22.6	45.6	23.2	48.2	25.3	69.6	22.0	56.3	10.0
3	48.8	18.1	27.7	23.2	44.8	22.7	49.0	25.7	70.3	21.7	56.9	10.2
4	46.9	17.8	27.6	23.3	45.4	23.2	49.1	26.1	69.0	20.8	56.3	10.2

Employing more learnable meta-transformation heads is shown to improve the performance of our approach on different datasets, which demonstrates the effectiveness of exploiting different contrastive subspaces for better skeleton representation learning. It is worth noting that the model performance tends to be stable when applying many more heads, as it could introduce more random perturbation into contrastive learning and help obtain more robust prototype estimation (see Sec. 5.2.2).

5.4.7.2 Effects of Embedding Sizes

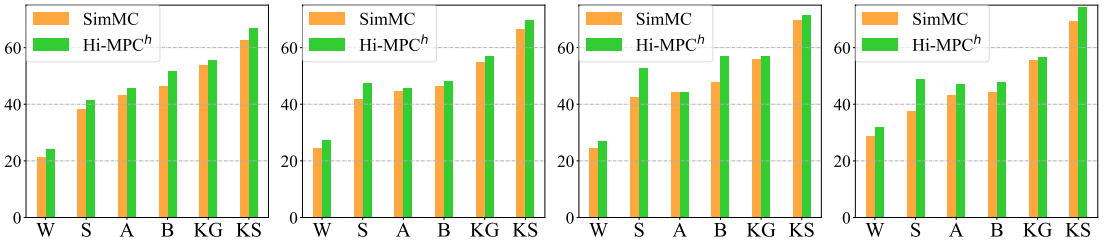
As shown in Table 5.11, our approach achieves higher performance with relatively larger embedding sizes ($h \geq 128$) on all datasets. The results suggest that a small embedding size ($h = 64$) could be insufficient to learn effective skeleton representations for person re-ID, while using too large sizes of embedding (*e.g.*, $h = 512$) might cause the model to learn more redundant feature information and degrade the overall performance.

5.4.7.3 Different Settings of DBSCAN

We evaluate the effects of the two main parameters in the DBSCAN algorithm, *i.e.*, minimum sample amount a_{min} within the maximum distance ϵ, which are empirically selected to encourage more balanced and stable clustering. As presented in Table 5.12, a_{min} seems to have no significant effect on the performance of

TABLE 5.13: Performance of our approach when setting different maximum distances ($\epsilon = 0.4, 0.6, 0.8, 1.0$) for DBSCAN.

ϵ	BIWI-S		BIWI-W		IAS-A		IAS-B		KS20		KGBD	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
0.4	46.8	16.6	19.9	18.7	43.2	18.2	42.0	20.0	69.5	20.7	55.6	7.8
0.6	47.5	17.4	22.4	19.2	44.7	20.7	46.7	25.8	69.4	21.8	56.9	10.2
0.8	50.3	18.2	27.3	22.6	45.6	23.2	48.2	25.3	69.6	22.0	47.9	4.8
1.0	27.9	9.7	11.6	14.2	33.8	13.6	37.2	14.3	50.0	11.0	46.9	4.3



(A) $F = 4$ (B) $F = 6$ (C) $F = 8$ (D) $F = 10$

FIGURE 5.12: Multi-shot person re-ID performance (Rank-1 accuracy) of our approach with different settings of sequence lengths ($F = 4, 6, 8, 10$) on BIWI-W (W), BIWI-S (S), IAS-A (A), IAS-B (B), KGBD (KG), and KS20 (KS) testing sets. The latest state-of-the-art method SimMC is compared as the performance baseline.

our approach, as setting different values of a_{min} achieve similar accuracy on most datasets. Nevertheless, it is worth mentioning that employing too small a value for a_{min} tends to generate much larger clusters and might lead to a degeneration of clustering (*e.g.*, single super cluster) and unstable model training in practice.

Large values of ϵ (*e.g.*, $\epsilon = 1.0$) greatly reduce the performance of our approach, as shown in Table 5.13. Considering that larger ϵ leads to higher connectedness of instances, *i.e.*, larger cluster and smaller number of prototypes, the results demonstrate that setting a relatively lower ϵ value with more diverse skeleton prototypes could facilitate learning richer discriminative features for person re-ID. However, too small a value for ϵ could cause excessive over-clustering, which leads to the degradation of model performance on some datasets.

5.4.7.4 Effects of Sequence Lengths

We evaluate multi-shot performance of our approach with different settings of sequence lengths F (*i.e.*, F-shot person re-ID), and compare it with the latest SimMC [5] on different datasets. As shown in Fig. 5.12, the proposed Hi-MPC^h consistently outperforms SimMC on all cases of datasets and sequence lengths. In contrast to

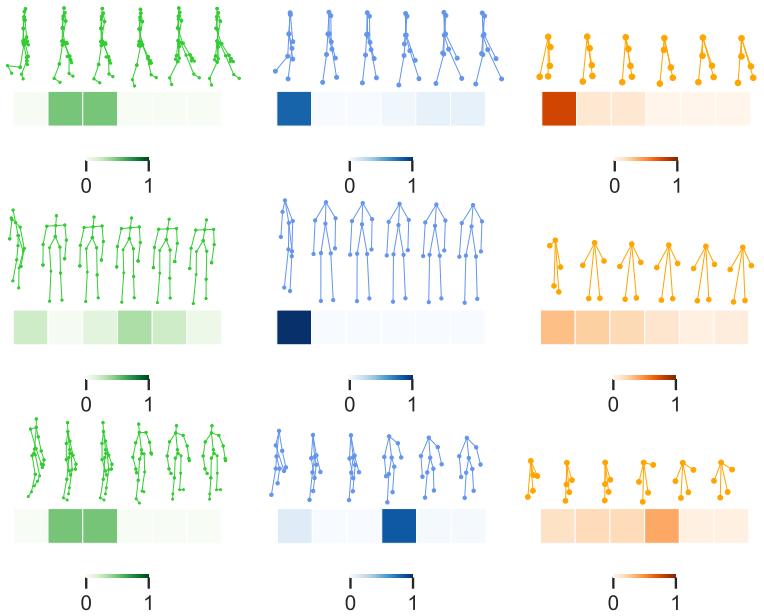


FIGURE 5.13: Visualization of joint-level (green), component-level (blue), limb-level representations (orange) of consecutive skeletons and their informative importance in different datasets. Each row shows three level representations of the same sequence. Darker colors of i^{th} position in heat maps indicate higher importance of i^{th} skeleton representation.

TABLE 5.14: Performance of our approach when applying heterogeneous (Heter.) or homogeneous (Homo.) feature mapping for transforming skeleton instances and prototypes. Note: We adopt homogeneous feature mapping in the proposed meta-transformation heads.

Types	BIWI-S		BIWI-W		IAS-A		IAS-B		KGBD		KS20	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
Heter.	37.2	10.8	15.0	16.2	38.7	15.4	42.2	16.4	49.9	6.3	64.5	15.7
Homo.	47.5	17.4	27.3	22.6	45.6	23.2	48.2	25.3	56.9	10.2	69.6	22.0

SimMC that fails to keep high accuracy when F varies slightly, our approach is more stable with better results on different datasets. Since skeleton sequences contain more pattern features as F increases, our approach is capable of learning more effective skeleton representations to achieve larger performance improvement in most cases. Nevertheless, it is interesting to note that using shorter sequences sometimes performs better than longer sequences on small datasets such as IAS-B, implying that a larger size of available training sequences under smaller F settings could help learn better representations on those datasets.

5.4.7.5 Effects of Different Types of Feature Mapping

We replace the homogeneous feature mapping (*i.e.*, $\mathbf{H}_1^{l,m} = \mathbf{H}_2^{l,m}$) used in meta-transformation heads with the heterogeneous feature mapping (*i.e.*, $\mathbf{H}_1^{l,m} \neq \mathbf{H}_2^{l,m}$), and evaluate the performance of our approach when keeping other configurations unchanged. As shown in Table 5.14, the model performance drops significantly on all datasets when applying heterogeneous (Heter.) feature mapping. This is because that the instances and prototypes that come from the same feature space are separately mapped into two different feature subspaces, where the inherent domain shifts [135] could hinder the proposed meta-prototype contrastive learning and hence reduce the model performance. As we expect the original instances and corresponding prototypes can be contrasted in the same new feature subspace, we employ homogeneous feature mapping in the proposed meta-transformation.

5.4.8 Analysis of Hard Skeleton Mining

To verify the effectiveness of the proposed HSM mechanism on mining more informative skeletons, we visualize joint-level, component-level, and limb-level skeleton representations with their inferred importance on different datasets. As shown in Fig. 5.13, the higher importance is often assigned to either evidently different skeletons or dramatically changing poses in the sequence, which could introduce more pattern information compared with other similar skeletons. This is consistent with our intuition that skeletons containing diverse patterns (*i.e.*, higher intra-class variation) are harder to be recognized as the same person, thus deserving more attention to learn during training. It is also observed that there exists a good alignment of informative importance between component-level and limb-level skeleton representations, while our model focuses on more key skeletons with continuous patterns at joint-level, which suggests that the proposed approach may hierarchically capture different skeleton semantics (*e.g.*, pattern consistency) to mine more effective features from various levels.

The proposed HSM mechanism allows us to intuitively visualize the importance/-value of each skeleton in learning *hard* patterns and useful features. As shown in Fig. 5.13, the component-level skeleton representations with a simpler body structure are highly similar with the joint-level skeleton representations to effectively characterize body poses, while the limb-level ones can provide more global

motion dynamics of skeletons. This further suggests the potential of more concise and abstract skeleton representations on learning unique patterns for person re-ID. Moreover, as hard skeletons typically contain easily-confused or uncommon patterns, we can potentially exploit hard skeleton mining to detect special skeletons/poses (*e.g.*, abnormal or pathological gaits) of a certain person for more advanced tasks such as medical gait analysis. It could also be used to discover noisy skeletons with incomplete or extremely unnatural poses for efficient skeleton filtration and selection.

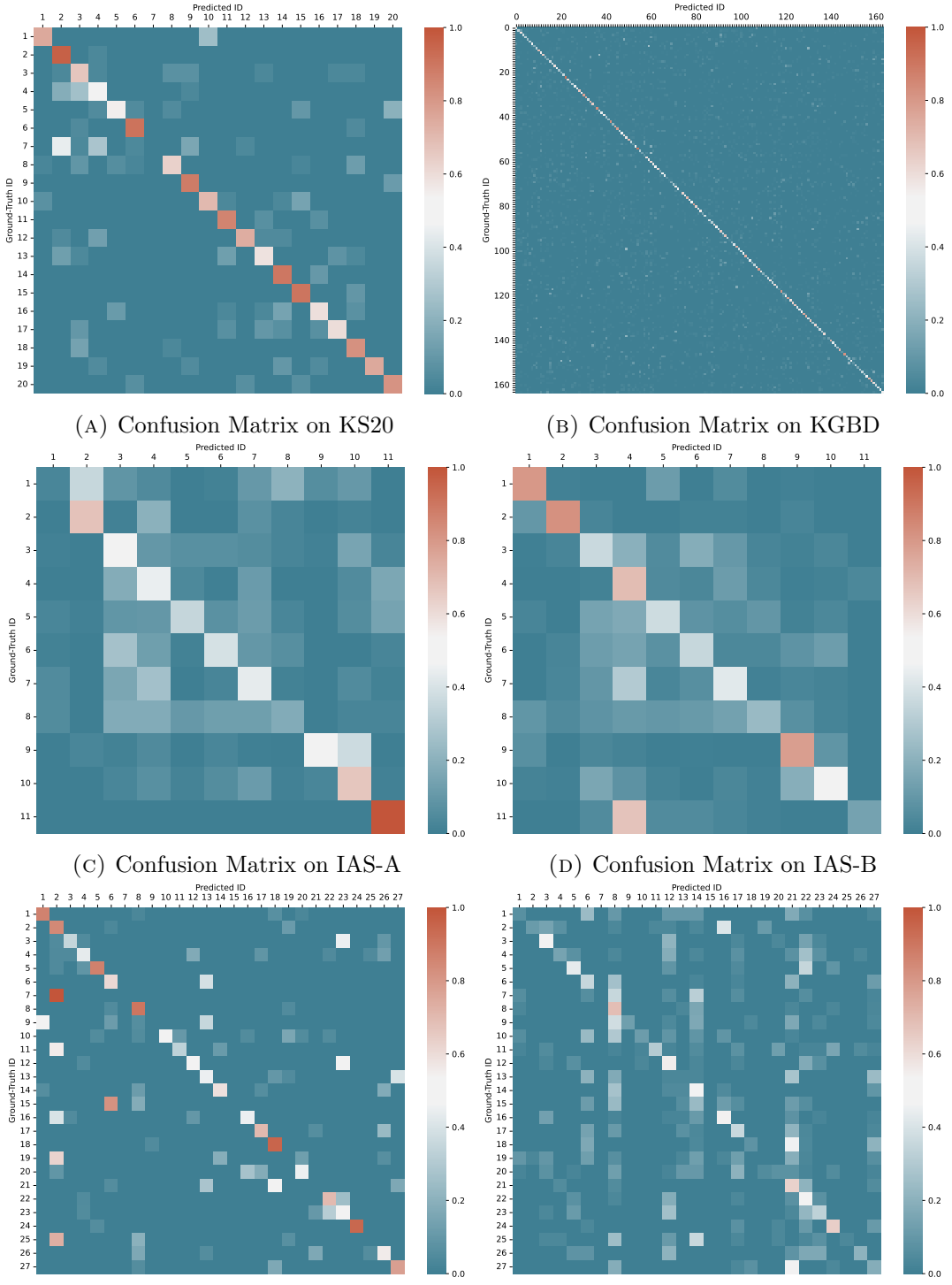
5.4.9 Analysis of Training Process

We visualize the total training loss ($\mathcal{L}_{\text{Hi-MPC}^h}$) in Fig. 5.8, which shows that the training of our approach can converge very fast in the first 100 optimization epochs. To provide a further analysis of the learned skeleton representations, we follow [129, 138] to estimate the *mutual information (MI)* and *adjusted mutual information score (AMI)* between the clustered skeleton features and the ground-truth identity labels, as shown in Fig. 5.9 and Fig. 5.10, respectively. The results show that the training of our approach rapidly and significantly improves both MI and AMI *w.r.t.* skeleton representations, *i.e.*, similarity between the pseudo classes generated by our approach and ground-truth class labels, which demonstrates that the proposed hierarchical skeleton meta-prototype contrastive learning can encourage the model to capture class-related semantics (*e.g.*, inter-class differences) to learn more discriminative skeleton representations.

5.4.10 Analysis of Confusion Matrix

In Fig. 5.14, we show the confusion matrices of our approach when performing person re-ID with the Rank-1 matching (*i.e.*, predicting the identity of each probe sequence using the top-1 gallery sequence that has the smallest Euclidean distance) on all testing sets (probe sets). As presented in Fig. 5.14 (a)-(f), each confusion matrix shows an evident alignment between the predicted identities and the ground-truth identities on the diagonal line. This suggests that skeleton sequences in most classes can be correctly matched between the probe and gallery set of each dataset. Moreover, it can be seen that the numbers of classes with high accuracy (*i.e.*,

number of red grids on the diagonal line) in KS20, KGBD, IAS-B and BIWI-Still are larger than that in IAS-A and BIWI-Walking. The larger numbers of white and red grids diffused around the diagonal lines, which represent the higher proportions of false matches, on the matrices of IAS-A (see Fig. 5.14 (c)) and BIWI-Walking (see Fig. 5.14 (f)) imply that our model tends to confuse skeleton sequences of more different identities on these datasets. This is consistent with the performance results shown in our work.



(A) Confusion Matrix on KS20

(B) Confusion Matrix on KGBD

(C) Confusion Matrix on IAS-A

(D) Confusion Matrix on IAS-B

(E) Confusion Matrix on BIWI-Still

(F) Confusion Matrix on BIWI-Walking

FIGURE 5.14: Visualization of confusion matrices on KS20 (a), KGBD (b), IAS-A (c), IAS-B (d), BIWI-Still (e), and BIWI-Walking (f) when using the Rank-1 matching.

5.5 Theoretical Hypotheses and Analyses

The proposed skeleton Meta-Prototype Contrastive (MPC) learning can be formulated as Expectation-Maximization (EM) solutions. In this section, we first provide a theoretical EM modeling for MPC to prove its validity and convergence, and then present hypotheses and analysis for the effectiveness of the proposed approach (Hi-MPC) that combines hierarchical skeleton representations and MPC.

Preliminaries. For clarity and convenience, we adopt a more general notation here, which is different from that used in the previous parts of Chapter 5. Suppose that a training set $X = \{\boldsymbol{x}_i\}_{i=1}^N$ contains N skeleton sequences, where $\boldsymbol{x}_i \in \mathbb{R}^{F \times S_n}$, F is the sequence length, and $n \in \{1, 2, 3\}$ denotes the level number of hierarchical skeleton representations[3]. The objective of unsupervised skeleton representation learning is to learn an embedding/encoder function f_θ (realized via θ-parameterized neural networks) that maps X to $V = \{\boldsymbol{v}_i\}_{i=1}^N$, where $\boldsymbol{v}_i \in \mathbb{R}^H$, by $\boldsymbol{v}_i = f_\theta(\boldsymbol{x}_i)$ without using any label, such that $\boldsymbol{v}_i$ can effectively represent latent features of $\boldsymbol{x}_i$ to perform person re-identification.

Formally, the goal is to find the network parameter θ that maximizes the log-likelihood function of the observed N skeleton sequences as follows:

$$\begin{aligned} \theta^* &= \arg\max_\theta L(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N; \theta) \\ &= \arg\max_\theta \prod_{i=1}^N p(\boldsymbol{x}_i; \theta) \\ &\iff \arg\max_\theta \sum_{i=1}^N \log p(\boldsymbol{x}_i; \theta), \end{aligned} \tag{1}$$

where $L(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N; \theta)$ denotes the likelihood function of the observed skeleton sequences $\{\boldsymbol{x}_i\}_{i=1}^N$ *w.r.t* θ, and each skeleton sequence $\boldsymbol{x}_i$ is hypothetically related to a certain skeleton prototype $\boldsymbol{c}_j \in \mathbb{R}^H$, with $\boldsymbol{c}_j \in \{\boldsymbol{c}_j\}_{j=1}^K$ and K is the number of prototypes. Under this assumption, we can re-formulate the objective in Eq. (1)

[3]The proposed hierarchical skeleton representations have $S_1 = 3 \times J$, $S_2 = 3 \times 10$, and $S_3 = 3 \times 5$ corresponding to joint-level, component-level, and limb-level representations. The EM modeling provided in this section is universal and can be independently applied to each level.

as:

$$\begin{aligned}\theta^* =& \arg\max_{\theta} \sum_{i=1}^{N} \log p\left(\boldsymbol{x}_i; \theta\right) \\ =& \arg\max_{\theta} \sum_{i=1}^{N} \log \sum_{j=1}^{K} p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right), \end{aligned} \tag{2}$$

Directly optimizing this function is intractable, thus we consider a lower-bound by using a surrogate function as:

$$\begin{aligned} & \sum_{i=1}^{N} \log \sum_{j=1}^{K} p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right) \\ =& \sum_{i=1}^{N} \log \sum_{j=1}^{K} Q\left(\boldsymbol{c}_j\right) \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right)}{Q\left(\boldsymbol{c}_j\right)} \\ \geq& \sum_{i=1}^{N} \sum_{j=1}^{K} Q\left(\boldsymbol{c}_j\right) \log \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right)}{Q\left(\boldsymbol{c}_j\right)}, \end{aligned} \tag{3}$$

where $Q\left(\boldsymbol{c}_j\right)$ represents some distribution over $\{\boldsymbol{c}_j\}_{j=1}^{K}$ and $\sum_{j=1}^{K} Q\left(\boldsymbol{c}_j\right) = 1$. We apply Jensen's inequality to derive the last step of Eq. (3), where equality can be achieved under the condition that $\frac{p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta)}{Q(\boldsymbol{c}_j)}$ is a constant. To realize this equality, we have:

$$\begin{aligned} Q\left(\boldsymbol{c}_j\right) = \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right)}{\sum_{m=1}^{K} p\left(\boldsymbol{x}_i, \boldsymbol{c}_m; \theta\right)} &= \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right)}{p\left(\boldsymbol{x}_i; \theta\right)} \\ &= p\left(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta\right), \end{aligned} \tag{4}$$

where $Q\left(\boldsymbol{c}_j\right)$ is a posterior probability related to $\boldsymbol{c}_j, \boldsymbol{x}_i$, and θ. When θ is fixed at the Expectation step, the distribution of representations $(\boldsymbol{x}_i)$ and corresponding prototypes $(\boldsymbol{c}_j)$ can be estimated as a result of clustering, thus we can get the constant value of $Q\left(\boldsymbol{c}_j\right) = p\left(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta\right)$ based on the result. We can re-write Eq. (3) as:

$$\sum_{i=1}^{N} \sum_{j=1}^{K} \left(Q\left(\boldsymbol{c}_j\right) \log p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right) - Q\left(\boldsymbol{c}_j\right) \log Q\left(\boldsymbol{c}_j\right)\right), \tag{5}$$

where the constant $-\sum_{i=1}^{N}\sum_{j=1}^{K} Q\left(\boldsymbol{c}_j\right)\log Q\left(\boldsymbol{c}_j\right)$ can be ignored and we need to maximize:

$$\sum_{i=1}^{N}\sum_{j=1}^{K} Q\left(\boldsymbol{c}_j\right)\log p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right). \tag{6}$$

For **the Expectation (E)-step**, we aim to estimate $p\left(\boldsymbol{c}_j;\boldsymbol{x}_i,\theta\right)$ (see Eq. (4)). In our approach, we run the DBSCAN algorithm on the encoded features $\boldsymbol{v}_i = f_\theta\left(\boldsymbol{x}_i\right)$ to obtain K clusters[4] $\{\boldsymbol{C}_j\}_{j=1}^{K}$. We generate corresponding skeleton prototype $\boldsymbol{c}_j$, which is the centroid of the j^{th} cluster $\boldsymbol{C}_j$. Then, we compute $p\left(\boldsymbol{c}_j;\boldsymbol{x}_i,\theta\right) = \mathbb{1}\left(\boldsymbol{x}_i \in \boldsymbol{C}_j\right)$, where $\mathbb{1}\left(\boldsymbol{x}_i \in \boldsymbol{C}_j\right) = 1$ if $\boldsymbol{x}_i$ belongs to the j^{th} cluster $\boldsymbol{C}_j$ (*i.e.*, corresponding to skeleton prototype $\boldsymbol{c}_j$); otherwise $\mathbb{1}\left(\boldsymbol{x}_i \in \boldsymbol{C}_j\right) = 0$.

Assumption 5.1. ***Prototype-Cluster Consistency.*** The global distribution of prototypes is consistent with the distribution of cluster centroids, *i.e.*, each cluster explicitly corresponds to the group of instances that belong to the same prototype. In the E-step, we adopt this commonly-used assumption [5, 88] to generate skeleton prototypes and derive $p\left(\boldsymbol{c}_j;\boldsymbol{x}_i,\theta\right) = \mathbb{1}\left(\boldsymbol{x}_i \in \boldsymbol{C}_j\right)$.

In the **Maximization (M)-step**, we combine Eq. (4) to maximize the lower-bound in Eq. (6) after the E-step:

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{j=1}^{K} Q\left(\boldsymbol{c}_j\right)\log p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right)\\
=&\sum_{i=1}^{N}\sum_{j=1}^{K} p\left(\boldsymbol{c}_j;\boldsymbol{x}_i,\theta\right)\log p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right)\\
=&\sum_{i=1}^{N}\sum_{j=1}^{K} \mathbb{1}\left(\boldsymbol{x}_i \in \boldsymbol{C}_j\right)\log p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right).
\end{aligned} \tag{7}$$

Each cluster centroid $\boldsymbol{c}_j$ is assumed to have a uniform prior probability $p\left(\boldsymbol{c}_j;\theta\right) = \frac{1}{K}$ since we are not provided any samples. We have:

$$\begin{aligned}
p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right) &= p\left(\boldsymbol{x}_i;\boldsymbol{c}_j,\theta\right)p\left(\boldsymbol{c}_j;\theta\right)\\
&= \frac{1}{K}\cdot p\left(\boldsymbol{x}_i;\boldsymbol{c}_j,\theta\right),
\end{aligned} \tag{8}$$

[4]The clusters $\{\boldsymbol{C}_j\}_{j=1}^{K}$ of $\boldsymbol{x}_i$ are generated based on the clustering of their encoded features $\boldsymbol{v}_i$.

where the distribution of samples around each prototype is assumed to be an isotropic Gaussian, leading to:

$$p\left(\boldsymbol{x}_i; \boldsymbol{c}_j, \theta\right) = \frac{\exp\left(\frac{-(\boldsymbol{v}_i - \boldsymbol{c}_p)^2}{2\sigma_p^2}\right)}{\sum_{j=1}^{K} \exp\left(\frac{-(\boldsymbol{v}_i - \boldsymbol{c}_j)^2}{2\sigma_j^2}\right)}, \tag{9}$$

where $\boldsymbol{v}_i = f_\theta\left(\boldsymbol{x}_i\right)$ and $\boldsymbol{c}_p$ is the prototype for the cluster $\boldsymbol{C}_p$ containing $\boldsymbol{x}_i$, *i.e.*, $\boldsymbol{x}_i \in \boldsymbol{C}_p$. We apply ℓ_2-normalization to both $\boldsymbol{v}$ and $\boldsymbol{c}$ to have $(\boldsymbol{v} - \boldsymbol{c})^2 = 2 - 2\boldsymbol{v} \cdot \boldsymbol{c}$. Then combining this with Eqs. (2), (3), (6), (7), (8), and (9), we can get the maximum log-likelihood estimation with:

$$\begin{aligned} \theta^* &= \arg\min_\theta \sum_{i=1}^{N} -\log \frac{\exp\left(\boldsymbol{v}_i \cdot \boldsymbol{c}_p / \tau_p\right)}{\sum_{j=1}^{K} \exp\left(\boldsymbol{v}_i \cdot \boldsymbol{c}_j / \tau_j\right)} \iff \\ \theta^* &= \arg\min_\theta \sum_{k=1}^{K} \sum_{i=1}^{N_k} -\log \frac{\exp\left(\boldsymbol{v}_i^k \cdot \boldsymbol{c}_k / \tau_k\right)}{\sum_{j=1}^{K} \exp\left(\boldsymbol{v}_i^k \cdot \boldsymbol{c}_j / \tau_j\right)}, \end{aligned} \tag{10}$$

where $\boldsymbol{v}_i^k$ denotes the representation of i^{th} sample (*i.e.*, skeleton instance) belonging to the k^{th} prototype $\boldsymbol{c}_k$, N_k is the number of samples in the k^{th} cluster, and τ is related to the distribution of features around different prototypes.

Assumption 5.2. ***Maximum Homogeneous Similarity.*** The homogeneous instances, which are defined as instances within the same cluster, should share higher inherent similarity than heterogeneous instances between different clusters. In other words, the prototype of each cluster can represent a unique skeleton concept or attribute of a certain identity, and the same cluster's instances possess the homogeneity of features corresponding to this prototype [131]. According to Assumption 5.1, it can be equivalent to the objective that each instance should be maximally similar to the corresponding prototype and be minimally similar to other prototypes. In the M-step, we maximize the probability that each instance belongs to its unique prototype (see Eq. (9)) based on this assumption. The equivalent formulation of this objective in Eq. (10) after applying feature ℓ_2-normalization can be further interpreted as to maximize the dot-product based similarity between instances and their prototypes while maximizing the dissimilarity to other prototypes.

Relations to Existing Contrastive Losses: (1) The InfoNCE loss [132] reformulated in MoCo [94] and SimCLR [91] can be interpreted as special cases of

the maximum log-likelihood estimation in Eq. (10), where the prototype $\boldsymbol{c}_p$ for a feature $\boldsymbol{v}_i$ is replaced by the augmented feature $\boldsymbol{v'}_i$ generated from different views of augmentation of the same instance (*i.e.*, $\boldsymbol{c}_p = \boldsymbol{v'}_i$) and τ is fixed as a temperature for contrastive learning. **(2)** The ProtoNCE loss used in PCL [88] has a similar form as Eq. (10), where τ is estimated with the assumption that the distribution of feature representations around each prototype varies in different clusters. However, PCL estimates the feature distribution under the Euclidean distance metric used in the k-means clustering. Such estimation could be inapplicable (*e.g.*, can not be generalized) to models that employ different clustering algorithms (*e.g.*, density-based DBSCAN [6]) or/and different distance metrics (*e.g.*, Jaccard metric), thus failing to getting satisfactory performance in practice [5].

Temperatures. In our work, we adopt a generic form following the common practice [81, 91, 94], *i.e.*, setting a global temperature τ for the proposed approach. By assuming a uniform feature distribution around each instance (*i.e.*, $\tau = \tau_k = \tau_j$), we encourage the model to learn representations with higher global uniformity, which could improve the quality of contrastive representation learning as theoretically and empirically proved in [5, 86, 130].

In the proposed approach, instead of directly using the original prototypes and instances, we transform them into meta-prototypes and meta-instances using different homogeneous transformation heads, so as to perform Meta-Prototype Contrastive learning (MPC) in different contrastive learning subspaces. According to the prototype-cluster consistency (see Assumption 5.1), we can exploit the root estimation of clusters and prototypes, *i.e.*, original contrastive space, to guide the contrast between meta-instances and meta-prototypes in different contrastive subspaces inheriting from the original space of prototypes, so as to encourage more robust probability estimation of prototypes and more consistent contrastive learning. We can formulate the proposed MPC loss based on Eq. (10) as:

$$\mathcal{L}_{\text{MPC}} = \sum_{k=1}^{K}\sum_{j=1}^{N_k}\sum_{m=1}^{M} -\log \frac{\exp\left((\hat{\boldsymbol{v}}_j^k)^m \cdot (\hat{\boldsymbol{c}}^k)^m/\tau\right)}{\sum_{u=1}^{K}\exp\left((\hat{\boldsymbol{v}}_j^k)^m \cdot (\hat{\boldsymbol{c}}^u)^m/\tau\right)}, \tag{11}$$

where $(\hat{\boldsymbol{v}}_j^k)^m = \mathbf{H}_1^m \boldsymbol{v}_j^k$ and $(\hat{\boldsymbol{c}}^k)^m = \mathbf{H}_2^m \boldsymbol{c}^k$ denote the m^{th} transformed meta-instance and meta-prototype corresponding to the original instance $\boldsymbol{v}_j^k$ and prototype $\boldsymbol{c}^k$, $\mathbf{H}_1^m, \mathbf{H}_2^m \in \mathbb{R}^{h_2 \times h_1}$ are the learnable weight matrices of the m^{th} meta-transformation head (we adopt $\mathbf{H}_1^m = \mathbf{H}_2^m$ and $h = h_1 = h_2$ for homogeneous transforming as detailed in Sec. 5.2), K denotes the number of skeleton prototypes generated from the original skeleton instances, N_k is the number of skeleton instances belonging to the k^{th} prototype $\boldsymbol{c}^k$ (equivalent to $\hat{\boldsymbol{p}}_+$ in Sec. 5.2), and τ represents the global temperature for contrastive learning. Note that $\mathcal{L}_{\text{MPC}}$ is averaged over all meta-instances for training. When we utilize one or zero transformation head for skeleton prototype learning, *i.e.*, $M = 1$ or 0, the objective of Eq. (11) is theoretically equivalent to Eq. (10), which is defined as direct prototype contrastive learning (DPC) in our work. The proposed MPC exploits the prototype-cluster consistency (see Assumption 5.1) to perform meta prototype-instance contrasting, which could enhance the inherent consistency and representativeness of learnable prototypes by jointly attending to key meta-prototypes in different representation subspaces. Intuitively, DPC is performed on a single representation space, which can be viewed as a special case of MPC.

Convergence Proof

We prove the convergence of MPC under modeling the maximum log-likelihood estimation (see Eq. (10)). Recall Eqs. (2) and (3) and let

$$\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{N} \log p\left(\boldsymbol{x}_i; \theta\right) \\
&= \sum_{i=1}^{N} \log \sum_{j=1}^{K} p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right) \\
&= \sum_{i=1}^{N} \log \sum_{j=1}^{K} Q\left(\boldsymbol{c}_j\right) \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right)}{Q\left(\boldsymbol{c}_j\right)} \\
&\geq \sum_{i=1}^{N} \sum_{j=1}^{K} Q\left(\boldsymbol{c}_j\right) \log \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta\right)}{Q\left(\boldsymbol{c}_j\right)}.
\end{aligned} \tag{12}$$

The above inequality holds with equality when $Q\left(\boldsymbol{c}_j\right) = p\left(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta\right)$ is a constant (see Eq. (4)).

In the t^{th} E-step, we have estimated the constant value $Q^{(t)}(\boldsymbol{c}_j) = p\left(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta^{(t)}\right)$. Then we have:

$$\ell\left(\theta^{(t)}\right) = \sum_{i=1}^{N}\sum_{j=1}^{K} Q^{(t)}(\boldsymbol{c}_j) \log \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta^{(t)}\right)}{Q^{(t)}(\boldsymbol{c}_j)}. \tag{13}$$

For the t^{th} M-step, we fix $Q^{(t)}(\boldsymbol{c}_j) = p\left(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta^{(t)}\right)$ and train model parameters θ to maximize Eq. (13). In this way, we can always have:

$$\begin{aligned}
\ell\left(\theta^{(t+1)}\right) &\geq \sum_{i=1}^{N}\sum_{j=1}^{K} Q^{(t)}(\boldsymbol{c}_j) \log \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta^{(t+1)}\right)}{Q^{(t)}(\boldsymbol{c}_j)} \\
&\geq \sum_{i=1}^{N}\sum_{j=1}^{K} Q^{(t)}(\boldsymbol{c}_j) \log \frac{p\left(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta^{(t)}\right)}{Q^{(t)}(\boldsymbol{c}_j)} \\
&= \ell\left(\theta^{(t)}\right).
\end{aligned} \tag{14}$$

The above result that $\ell\left(\theta^{(t)}\right)$ monotonically increases with more iterations suggests the convergence of the algorithm.

The detailed convergence properties of the EM algorithm are discussed in [139, 140]. Here we only discuss the general case, and follow [140] to make the assumptions for the EM algorithm:

(a) Ω is a subset in the r-dimensional Euclidean space $\mathbb{R}^r$,
(b) $\Omega_{\theta^{(0)}} = \left\{\theta \in \Omega : \ell(\theta) \geq \ell\left(\theta^{(0)}\right)\right\}$ is compact for any $\ell\left(\theta^{(0)}\right) > -\infty$,
(c) $\ell(\,\cdot\,)$ is continuous in Ω and differentiable in the interior of Ω.

Under the assumptions of **(a)**, **(b)**, and **(c)**[5], we have:
(d) $\left\{\ell\left(\theta^{(t)}\right)\right\}_{t\geq 0}$ is bounded above for any $\theta^{(0)} \in \Omega$.

As a consequence of **(d)** and the inequality (14) (*i.e.*, $\ell\left(\theta^{(t+1)}\right) \geq \ell\left(\theta^{(t)}\right)$), $\ell\left(\theta^{(t)}\right)$ converges monotonically to some ℓ^*.

It is worth noting that there is no guarantee that ℓ^* is the global maximum of $\ell(\,\cdot\,)$ over Ω. As reported in previous works [140–144], if the log-likelihood function $\ell(\,\cdot\,)$ has several (local or global) maxima and stationary points, the convergence of the EM sequence $\left\{\ell\left(\theta^{(t)}\right)\right\}$ to either type of point depends on the choice of starting

[5]These assumptions can be satisfied in most practical situations. As the related proofs/discussions are out of the scope of this work, readers can refer to [140] for more details.

point. Readers can refer to [139, 140] for more details about different convergence cases of the EM algorithm.

The aforementioned fact may account for the performance changes (*i.e.*, small variations) of our model on the same dataset, as different random initializations of model parameters could change $\theta^{(0)}$ (*i.e.*, the starting point) hence the final convergence result. In practice, we follow [4, 5] to train the model with different random initializations on each dataset and report its average performance, which helps estimate a more stable EM convergence result with different initialized starting points.

Empirical Analysis

We empirically provide several interpretations for the effectiveness of MPC: **(1)** The proposed MPC exploits the prototype-cluster consistency (see Assumption 5.1) to perform meta prototype-instance contrasting. Compared with DPC, it can jointly attend to key meta-prototypes in different representation subspaces to learn more consistent and representative prototypes, which facilitates the model to capture more effective skeleton representations for person re-ID. **(2)** MPC can be viewed as adding random perturbation, *i.e.*, different initialized contrastive representations in new feature spaces, to the original direct skeleton prototype contrasting (DPC). The average integration of all losses from different meta-prototypes enables the model to obtain a more robust estimation of prototype loss[6] to enhance the contrastive learning of most typical skeleton features. **(3)** The extra linear transformation heads (*i.e.*, $M > 1$) can be interpreted as random *homogeneous* augmentations of skeleton features. Similar to the random data augmentation applied to the input data [91, 94], MPC adopts multiple randomly initialized transformation heads to augment intermediate features of the same skeleton sequence from different "views" for better contrastive learning and skeleton representation learning.

[6]PCL [88] demonstrates that a more robust probability estimation of prototypes can enhance the structure encoding and contrastive learning.

Coarse-to-Fine MPC: Hierarchical Meta-prototype Contrastive Learning (Hi-MPC)

The proposed Hi-MPC combines MPC of different level skeleton representations, as formulated with:

$$\mathcal{L}_{\text{Hi-MPC}} = \sum_{l=1}^{3}\sum_{k=1}^{K}\sum_{j=1}^{N_k}\sum_{m=1}^{M} -\log \frac{\exp\Big((\hat{\boldsymbol{v}}_j^{l,k})^m \cdot (\hat{\boldsymbol{c}}^{l,k})^m/\tau\Big)}{\sum_{u=1}^{K}\exp\Big((\hat{\boldsymbol{v}}_j^{l,k})^m \cdot (\hat{\boldsymbol{c}}^{l,u})^m/\tau\Big)}, \tag{15}$$

where $(\hat{\boldsymbol{v}}_j^{l,k})^m$ and $(\hat{\boldsymbol{c}}^{l,k})^m$ denote the m^{th} transformed meta-instance and meta-prototype of the l^{th} level skeleton representation belonging to the k^{th} cluster. Other notations are the same as Eq. (11). $\mathcal{L}_{\text{Hi-MPC}}$ can be re-formulated as:

$$\mathcal{L}_{\text{Hi-MPC}} = \sum_{l=1}^{3}\sum_{i=1}^{I_l}\sum_{m=1}^{M} -\log \frac{\exp\Big((\hat{\boldsymbol{v}}^{l,(i)})^m \cdot (\hat{\boldsymbol{c}}_{+}^{l})^m/\tau\Big)}{\sum_{k=1}^{K}\exp\Big((\hat{\boldsymbol{v}}^{l,(i)})^m \cdot (\hat{\boldsymbol{c}}_{k}^{l})^m/\tau\Big)}, \tag{16}$$

where I_l denotes the number of instances in all clusters generated from the l^{th} level skeleton representations, $(\hat{\boldsymbol{v}}^{l,(i)})^m$ denotes the m^{th} transformed meta-instance of i^{th} instance, $(\hat{\boldsymbol{c}}_{+}^{l})^m$ (equivalent to $(\hat{\boldsymbol{p}}_{+}^{l})^m$ in Sec. 5.2) is its corresponding meta-prototype, and $(\hat{\boldsymbol{c}}_{k}^{l})^m$ denotes the meta-prototype of the k^{th} cluster at the l^{th} level. The proposed Hi-MPC could be viewed as performing coarse-to-fine meta-prototype contrastive learning with hierarchical skeleton representations (*i.e.*, MPC using single-level skeleton representations is a special case of Hi-MPC), and allows the model to mine both low-level and high-level skeleton semantics (*e.g.*, identity-specific patterns, intra-class similarity) from different representation subspaces of skeleton prototype learning, which facilitates learning more discriminative skeleton representations for person re-ID, as demonstrated in our work.

5.6 Summary

In this chapter, we propose a hierarchical skeleton meta-prototype contrastive learning (Hi-MPC) approach with a hard skeleton mining (HSM) mechanism to

contrast and learn the most representative features from key informative skeletons for unsupervised person re-ID. The hierarchical representations of 3D skeletons are built to capture coarse-to-fine body features from different levels. To encourage more consistent contrastive learning to mine more representative skeleton prototypes, we propose to perform meta-transformation of instances and prototypes to contrast in multiple contrastive feature subspaces. We further devise a hard skeleton mining mechanism to assign larger importance to harder skeletons, so as to learn more valuable patterns and effective representations. Finally, we combine different level skeleton features learned from our approach to construct multi-level skeleton meta-representation (MSMR) for person re-ID. Our approach outperforms most state-of-the-art methods, and is also highly effective when applied to multi-view and RGB-based scenarios with estimated skeletons.

Chapter 6

Skeleton-Based Person Re-ID with Body-Joint Relation Learning

6.1 Introduction

Existing methods typically extract anthropometric descriptors and gait attributes from body-joint coordinates [2, 24, 25], or leverage sequence learning paradigms such as Long Short-Term Memory (LSTM) [36] to model body and motion features with skeleton sequences [4, 69]. Some recent works explore new paradigms such as hierarchical cluster-contrast architectures [133] (see Chapter 5) and efficient Siamese architectures [5] (see Chapter 4) to enhance skeleton representation learning with coarse-to-fine motion modeling and high-level semantics learning. However, these methods rarely explore the inherent *body relations* within skeletons (*e.g.*, inter-joint motion correlations), thus largely ignoring some valuable skeleton patterns (corresponding to the third challenge in Sec. 1.2). To fill this gap, some previous studies [7, 61] construct skeleton graphs to model body-component relations in terms of structure and action. These methods typically require multi-stage

This chapter has been published as: Haocong Rao and Chunyan Miao, "TranSG: Transformer-Based Skeleton Graph Prototype Contrastive Learning with Structure-Trajectory Prompted Reconstruction for Person Re-Identification," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023 [14]. DOI: 10.1109/CVPR52729.2023.02118.

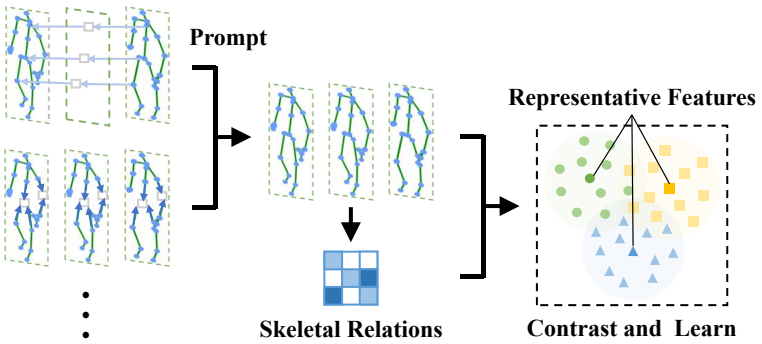


FIGURE 6.1: TranSG temporally and spatially masks skeleton graphs to prompt their reconstruction, while integrating full skeletal relations into contrastive learning of typical features for person re-ID.

non-parallel relation modeling, *e.g.*, [7] models collaborative relations conditioned on the structural relations, while they cannot simultaneously mine different underlying relations. On the other hand, they usually leverage *sequence-level* instances such as averaged features of sequential graphs [61] for representation learning, which inherently limits their ability to exploit richer graph semantics from fine-grained (*e.g.*, node-level) representations. For example, there usually exist strong correlations between nearby body-joint nodes within a local spatial-temporal graph context, which can *prompt* learning unique and recognizable skeleton patterns for person re-ID.

To address the above challenges, we propose a general Transformer-based Skeleton Graph prototype contrastive learning (TranSG) paradigm with structure-trajectory prompted reconstruction as shown in Fig. 6.1, which integrates different relational features of skeleton graphs and contrasts representative graph features to learn discriminative representations for person re-ID[1]. Specifically, we first model 3D skeletons as graphs, and propose the *Skeleton Graph Transformer (SGT)* to perform *full-relation* learning of body-joint nodes, so as to simultaneously aggregate key relational features of body structure and motion into effective graph representations. Second, a *Graph Prototype Contrastive learning (GPC)* approach is proposed to contrast and learn the most representative skeleton graph features (defined as *"graph prototypes"*) of each identity. By pulling together both *sequence-level* and *skeleton-level* graph representations to their corresponding prototypes while pushing apart representations of different prototypes, we encourage the model to capture discriminative graph features and high-level skeleton semantics (*e.g.*,

[1]Our codes are publicly available at https://github.com/Kali-Hac/TranSG.

identity-associated patterns) for person re-ID. Last, motivated by the inherent structural correlations and pattern continuity of body joints [4], we devise a *graph Structure-Trajectory Prompted Reconstruction (STPR) mechanism* to exploit the spatial-temporal context (*i.e.*, graph structure and trajectory) of skeleton graphs to prompt the skeleton graph reconstruction, which facilitates learning richer graph semantics and more effective graph representations for the person re-ID task.

With this chapter, we make the following contributions:

- We present a generic TranSG paradigm to learn effective representations from skeleton graphs for person re-ID. To the best of our knowledge, TranSG is the first *transformer* paradigm that unifies skeletal relation learning and skeleton graph contrastive learning specifically for skeleton-based person re-ID.
- We devise a skeleton graph transformer (SGT) to fully capture relations in skeleton graphs and integrate key correlative node features into graph representations.
- We propose the graph prototype contrastive learning (GPC) to contrast and learn the most representative graph features and identity-related semantics from both skeleton and sequence levels for person re-ID.
- We devise the graph structure-trajectory prompted reconstruction (STPR) to exploit spatial-temporal graph contexts for reconstruction, so as to capture more key graph semantics and unique features for person re-ID.

Extensive experiments on five public benchmarks show that TranSG prominently outperforms existing state-of-the-art methods and is highly scalable to be applied to different graph modeling, RGB-estimated or unlabeled skeleton data.

6.2 The Proposed TranSG Paradigm

Given a 3D skeleton sequence $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_f) \in \mathbb{R}^{f \times J \times 3}$, where $\boldsymbol{x}_t \in \mathbb{R}^{J \times 3}$ denotes the t^{th} skeleton with 3D coordinates of J body joints. Each skeleton sequence $\boldsymbol{X}$ corresponds to a person identity y, where $\mathrm{y} \in \{1, \cdots, \mathrm{C}\}$ and C is the number of different classes (*i.e.*, identities). We use $\Phi_T = \left\{\boldsymbol{X}_i^T\right\}_{i=1}^{N_1}$, $\Phi_P =$

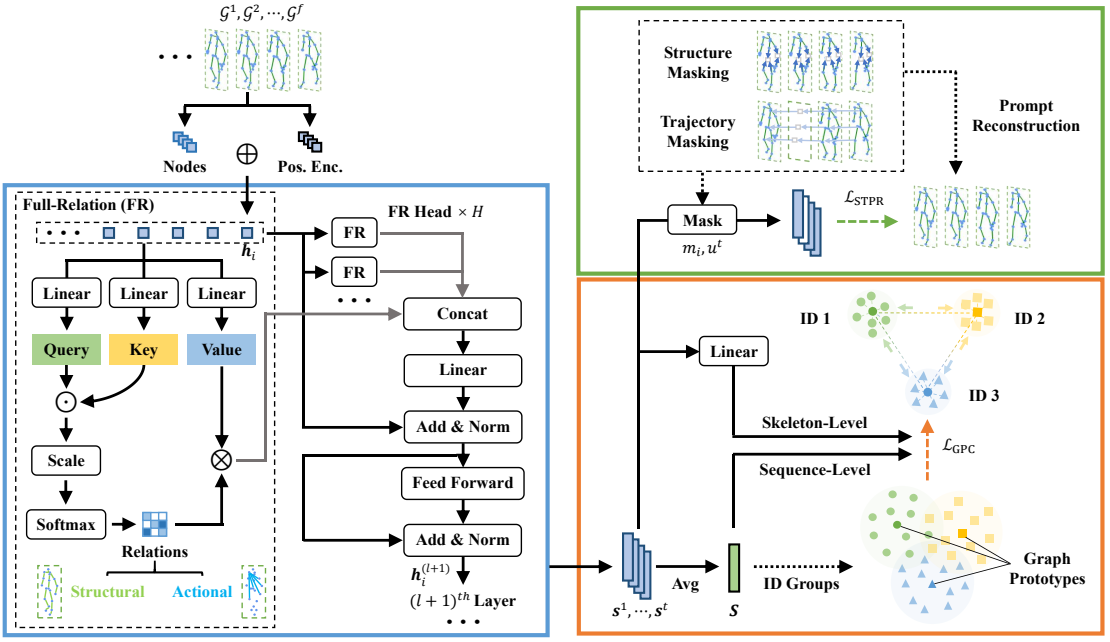


FIGURE 6.2: Schematics of TranSG approach with the skeleton graph transformer (SGT) (shown in the blue box), graph prototype contrastive learning (GPC) (presented in the orange box), and graph structure-trajectory prompted reconstruction (STPR) (presented in the green box).

$\{\boldsymbol{X}_i^P\}_{i=1}^{N_2}$, and $\Phi_G = \{\boldsymbol{X}_i^G\}_{i=1}^{N_3}$ to denote the *training* set, *probe* set, and *gallery* set that contain N_1, N_2, and N_3 skeleton sequences of different persons in different scenes and views. Our aim is to learn an encoder to map skeleton sequences into effective representations, such that the encoded skeleton sequence representations (denoted as $\{\boldsymbol{S}_i^P\}_{i=1}^{N_2}$) in the probe set can be matched with the representations (denoted as $\{\boldsymbol{S}_i^G\}_{i=1}^{N_3}$) of the same identity in the gallery set.

The overview of our TranSG approach is shown in Fig. 6.2: Firstly, we represent each skeleton sequence $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_f$ as skeleton graphs $\mathcal{G}^1, \cdots, \mathcal{G}^f$ (see Sec. 6.2.1), and integrate the graph positional information into their node representations $\boldsymbol{h}_i$. Then, they are fed into the skeleton graph transformer (SGT) to fully capture body-component relations with multiple full-relation (FR) heads (see Sec. 6.2.2). We employ multiple SGT layers and take the l^{th} layer output $\boldsymbol{h}_i^{(l+1)}$ as the input of $(l+1)^{th}$ layer. Next, the centroids of graph features belonging to different identities (ID) are utilized to generate *graph prototypes*, and we enhance the similarity of both skeleton-level ($\boldsymbol{s}^t$) and sequence-level graph representations ($\boldsymbol{S}$) to their corresponding prototypes, while maximizing their dissimilarity to other prototypes by optimizing $\mathcal{L}_{\text{GPC}}$ (see Sec. 6.2.3). Meanwhile, we randomly mask skeleton graph

structure and node trajectory, and exploit correspondingly masked graph representations as contexts to prompt reconstruction by minimizing $\mathcal{L}_{\text{STPR}}$ (see Sec. 6.2.4). The skeleton graph representations learned from our approach are exploited for person re-ID (see Sec. 6.2.5).

6.2.1 Skeleton Graph Construction

The human body with joints can be naturally modeled as graphs to characterize rich structural and positional information [8]. We construct skeleton graphs with the body joints as nodes and the structural connections between adjacent joints as edges. Each graph $\mathcal{G}^t(\mathcal{V}^t, \mathcal{E}^t)$ (corresponding to the t^{th} skeleton $\boldsymbol{x}_t$) consists of nodes $\mathcal{V}^t = \{\boldsymbol{v}_1^t, \boldsymbol{v}_2^t, \cdots, \boldsymbol{v}_J^t\}$, $\boldsymbol{v}_i^t \in \mathbb{R}^3$, $i \in \{1, \cdots, J\}$ and edges $\mathcal{E}^t = \{e_{i,j}^t \mid \boldsymbol{v}_i^t, \boldsymbol{v}_j^t \in \mathcal{V}^t\}$, $e_{i,j}^t \in \mathbb{R}$, where t is the index of skeleton in the sequence. Here $\mathcal{V}^t$ and $\mathcal{E}^t$ denote the set of nodes corresponding to J different body joints and the set of their internal connection relations, respectively. The adjacency matrix of $\mathcal{G}^t$ is denoted as $\mathbf{A}^t \in \mathbb{R}^{J \times J}$ to represent the relations among J nodes. $\mathbf{A}^t$ is initialized based on the connections of adjacent body joints, which is *learnable* during model training to capture dynamic relations and valuable skeleton patterns.

6.2.2 Skeleton Graph Transformer

As our goal is to capture discriminative skeleton features for person re-ID, it is crucial to consider two *unique* properties of human skeletons: (1) Body *structural* features, which can be inferred from the relations between adjacent body joints; (2) Skeleton *actional* patterns (*e.g.*, gait [37]), which are typically characterized by the relations among different body components [8]. From the perspective of graphs, we regard each body-joint node as a basic body component, and propose to combine the above relation learning as a *full-relation* learning of body-joint nodes, so as to fully aggregate key body and motion features from skeleton graphs. For this purpose, we devise the *skeleton graph transformer (SGT)* as follows (shown in Fig. 6.2).

Given a skeleton graph $\mathcal{G}^t(\mathcal{V}^t, \mathcal{E}^t)$ and its adjacency matrix $\mathbf{A}^t$, we first exploit the pre-defined graph structure to generate the positional encoding for graph nodes

with:

$$\Delta = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} = \mathbf{U}^T\mathbf{\Lambda}\mathbf{U}, \tag{6.1}$$

where $\mathbf{A}, \mathbf{D}$ are the adjacency matrix and degree matrix of the skeleton graph $\mathcal{G}$, respectively, and $\mathbf{\Lambda}, \mathbf{U}$ denote the matrices of Laplacian eigenvalues and eigenvectors, respectively. $\mathbf{U}^T\mathbf{\Lambda}\mathbf{U}$ is the factorization of the graph Laplacian matrix. For convenience, we use $\mathbf{A}$ to represent $\mathbf{A}^t$ as skeleton graphs in the same dataset share the identical initialized adjacency matrix. We follow [145] to adopt the K smallest non-trivial eigenvectors as the node positional encoding, denoted as $\boldsymbol{\lambda}_i \in \mathbb{R}^K$ for the node $\boldsymbol{v}_i$. They are mapped into feature spaces of the same dimension d with the affine transformation, which are then added by:

$$\boldsymbol{h}_i = (\mathbf{W}_v\boldsymbol{v}_i + \boldsymbol{b}_v) + (\mathbf{W}_p\boldsymbol{\lambda}_i + \boldsymbol{b}_p), \tag{6.2}$$

where $\boldsymbol{h}_i \in \mathbb{R}^d$ denotes the i^{th} position-encoded node representation, $\mathbf{W}_v \in \mathbb{R}^{d\times 3}, \mathbf{W}_p \in \mathbb{R}^{d\times K}, \boldsymbol{b}_v, \boldsymbol{b}_p \in \mathbb{R}^d$ are the learnable parameters of the feature transformation for the i^{th} node $\boldsymbol{v}_i$ and its corresponding positional encoding. Intuitively, the addition of positional encoding in Eq. (6.2) helps preserve the unique positional information of nodes based on the graph structure, *i.e.*, structurally nearby nodes are endowed with similar positional features while the farther nodes possess more dissimilar positional features, so as to encourage more effective node representation learning [145].

Then, given the body-joint node representations, we capture their inherent relations using multiple *independent* full-relation (FR) heads (see Fig. 6.2), and update node representations by aggregating corresponding relational features:

$$\boldsymbol{w}_{i,j}^{k,l} = \mathrm{Softmax}_j\left(\frac{(\boldsymbol{Q}^{k,l}\boldsymbol{h}_i^{(l)})\cdot(\boldsymbol{K}^{k,l}\boldsymbol{h}_j^{(l)})}{\sqrt{d_\mathrm{k}}}\right), \tag{6.3}$$

$$\hat{\boldsymbol{h}}_i^{(l)} = \boldsymbol{O}^l \Big\|_{k=1}^{H}\left(\sum_{j=1}^{J}\boldsymbol{w}_{i,j}^{k,l}\boldsymbol{V}^{k,l}\boldsymbol{h}_j^{(l)}\right), \tag{6.4}$$

where $\boldsymbol{Q}^{k,l}, \boldsymbol{K}^{k,l}, \boldsymbol{V}^{k,l} \in \mathbb{R}^{d_\mathrm{k}\times d}$ are the parameter matrices for query, key, and value transformations in the k^{th} FR head of the l^{th} SGT layer, $\boldsymbol{O}^l \in \mathbb{R}^{d\times d}$ is the parameter matrix for output transformation of the l^{th} SGT layer. $\frac{1}{\sqrt{d_\mathrm{k}}}$ is the scaling factor for the scaled dot-product similarity, $\boldsymbol{w}_{i,j}^{k,l}$ denotes the normalized relation value between the i^{th} and j^{th} node captured by the k^{th} FR head in the l^{th} layer, $\|$

represents the concatenation operation, and H is the number of FR heads. For clarity, we use $\hat{\boldsymbol{h}}_i^{(l)} \in \mathbb{R}^d$ to denote the i^{th} node representation that concatenates node features learned from different heads in the l^{th} layer. It is worth noting that SGT naturally generalizes the *self-attention* [108] to the full-relation learning of graph nodes, and can be viewed as a general paradigm that *simultaneously* captures structural and actional relations from both adjacent and non-adjacent body-component nodes. The multiple FR heads enable the model to *jointly* attend to node relations from different feature subspaces and integrate more key correlative node features into final node representations. We follow [145] to apply a Feed Forward Network (FFN) with residual connections [146] and batch normalization [147] by:

$$\overline{\boldsymbol{h}}_i^{(l)} = \text{Norm}\left(\boldsymbol{h}_i^{(l)} + \hat{\boldsymbol{h}}_i^{(l)}\right), \tag{6.5}$$

$$\boldsymbol{h}_i^{(l+1)} = \text{Norm}\left(\overline{\boldsymbol{h}}_i^{(l)} + \mathbf{W}_2^l\, \sigma\left(\mathbf{W}_1^l \overline{\boldsymbol{h}}_i^{(l)}\right)\right). \tag{6.6}$$

In Eqs. (6.5) and (6.6), Norm$(\cdot)$ denotes the batch normalization operation, $\mathbf{W}_1^l \in \mathbb{R}^{2d\times d}, \mathbf{W}_2^l \in \mathbb{R}^{d\times 2d}$ are the learnable parameters of FFN, $\sigma(\cdot)$ is the ReLU activation function, $\overline{\boldsymbol{h}}_i^{(l)}$ and $\boldsymbol{h}_i^{(l+1)}$ represent the intermediate and *output* node representations of the l^{th} SGT layer, respectively. We average the node features in each skeleton graph as the corresponding graph representation, and then integrate f consecutive graph representations into the final sequence-level graph representation $\boldsymbol{S}$ with:

$$\boldsymbol{S} = \frac{1}{f}\sum_{t=1}^{f} \boldsymbol{s}^t = \frac{1}{f}\sum_{t=1}^{f}\frac{1}{J}\sum_{i=1}^{J} \boldsymbol{h}_i^t, \tag{6.7}$$

where $\boldsymbol{S}, \boldsymbol{s}^t \in \mathbb{R}^d$ are the *sequence-level* and *skeleton-level* graph representations, corresponding to a skeleton sequence $\boldsymbol{X}$ and the t^{th} skeleton $\boldsymbol{x}_t$ in the sequence, respectively. For simplicity of presentation, we use $\boldsymbol{h}_i^t$ to denote the encoded representation of i^{th} node in the t^{th} skeleton graph. Here we assume that each node representation with aggregated relational features (see Eqs. (6.3), (6.4)) contributes equally to the graph representation, and each skeleton graph has the same importance in representing patterns of an individual.

6.2.3 Graph Prototype Contrastive Learning

Each individual's skeletons usually share the same anthropometric features (*e.g.*, skeletal lengths), while their continuous sequence can characterize identity-specific

walking patterns (*i.e.*, gait) [5]. In this context, it is desirable to mine the most representative skeleton features (defined as *"prototypes"*) of each individual to learn distinguishable patterns. A straightforward way is to cluster sequence representations to mine prototypes for contrastive learning like [5, 61], while they can only generate identity-agnostic (*i.e.*, pseudo-labeled) prototypes with large uncertainty or unreliability, *e.g.*, when existing large intra-class variation, two same-class sequences with diverse patterns might be clustered to two prototype groups. To encourage the model to generate representatives graph features more reliably, we propose to exploit the graph feature centroid of each *ground-truth* identity as *graph prototypes*, which are contrasted with both sequence-level and skeleton-level graph features to learn discriminative representations for person re-ID.

Given the encoded graph representations $\{\boldsymbol{S}_i^T\}_{i=1}^{N_1}$ of training skeleton sequences $\{\boldsymbol{X}_i^T\}_{i=1}^{N_1}$, we group them by ground-truth classes as $\{\widehat{\mathbb{S}}_k\}_{k=1}^{C}$, where $\widehat{\mathbb{S}}_k = \{\boldsymbol{S}_{k,j}\}_{j=1}^{n_k}$ denotes the set of graph representations belonging to the k^{th} identity, $\boldsymbol{S}_{k,j}$ is the j^{th} sequence-level graph representation, and n_k represents the number of k-class sequences. Then, the graph prototype is generated by averaging the graph features of the same class with:

$$\boldsymbol{c}_k = \frac{1}{n_k}\sum_{j=1}^{n_k}\boldsymbol{S}_{k,j}\ , \tag{6.8}$$

where $\boldsymbol{c}_k \in \mathbb{R}^d$ denotes the graph prototype of the k^{th} identity. To focus on the representative graph prototype of each identity to learn discriminative identity-related semantics from both skeleton and sequence levels, we propose the graph prototype contrastive (GPC) loss as:

$$\mathcal{L}_{\text{GPC}} = \alpha\mathcal{L}_{\text{GPC}}^{seq} + (1-\alpha)\mathcal{L}_{\text{GPC}}^{ske}, \tag{6.9}$$

where

$$\mathcal{L}_{\text{GPC}}^{seq} = \frac{1}{N_1}\sum_{k=1}^{C}\sum_{j=1}^{n_k} -\log\frac{\exp\left(\boldsymbol{S}_{k,j}\cdot\boldsymbol{c}_k/\tau_1\right)}{\sum_{m=1}^{C}\exp\left(\boldsymbol{S}_{k,j}\cdot\boldsymbol{c}_m/\tau_1\right)}, \tag{6.10}$$

$$\mathcal{L}_{\text{GPC}}^{ske} = \frac{1}{fN_1}\sum_{k=1}^{C}\sum_{j=1}^{n_k}\sum_{t=1}^{f} -\log\frac{\exp\left(\mathcal{F}_1\left(\boldsymbol{s}_{k,j}^t\right)\cdot\mathcal{F}_2\left(\boldsymbol{c}_k\right)/\tau_2\right)}{\sum_{m=1}^{C}\exp\left(\mathcal{F}_1\left(\boldsymbol{s}_{k,j}^t\right)\cdot\mathcal{F}_2\left(\boldsymbol{c}_m\right)/\tau_2\right)}. \tag{6.11}$$

In Eq. (6.9), α is the weight coefficient to fuse sequence-level ($\mathcal{L}_{\text{GPC}}^{seq}$) and skeleton-level graph prototype contrastive learning ($\mathcal{L}_{\text{GPC}}^{ske}$). In Eqs. (6.10) and (6.11), $\boldsymbol{c}_m$ represents the m-class graph prototype, $\boldsymbol{s}_{k,j}^t$ denotes the graph representation of the t^{th} skeleton in the sequence that corresponds to $\boldsymbol{S}_{k,j}$ belonging to the k^{th}

identity, τ_1, τ_2 are the temperatures for contrastive learning, and $\mathcal{F}_1(\cdot)$, $\mathcal{F}_2(\cdot)$ are linear projection heads to transform skeleton-level graph representations and graph prototypes into the same contrastive space $\mathbb{R}^d$. It should be noted that the graph prototypes are generated from higher level (*i.e.*, sequence-level) representations and the learnable linear projection in Eq. (6.11) can be viewed as integrating related graph features from both levels for contrastive learning. The proposed GPC loss is essentially a generalized skeleton prototype contrastive loss that combines joint-level relation encoding (see Sec. 6.2.2), skeleton-level and sequence-level graph learning (see Eq. (6.9)). Its objective can be theoretically modeled as an Expectation-Maximization (EM) solution [14].

6.2.4 Graph Structure-Trajectory Prompted Reconstruction

To exploit more valuable graph features and high-level semantics (*e.g.*, pattern consistency) from both spatial and temporal contexts of skeleton graphs, we propose a self-supervised *graph Structure and Trajectory Prompted Reconstruction (STPR)* mechanism. Motivated by the structural correlations and local motion continuity [4] of body components, we devise two graph context based prompts (see Fig. 6.2), namely (1) partial node positions of the *same* graph and (2) partial node trajectory among *continuous* graphs, to reconstruct the graph structure and dynamics.

Graph Structure Prompted Reconstruction. Given the t^{th} skeleton graph representation with J encoded nodes $(\boldsymbol{h}_1^t, \cdots, \boldsymbol{h}_J^t)$ encoded by SGT (see Eqs. (6.1)-(6.6)), we first randomly mask node positions to generate the masked graph representation as:

$$\hat{\boldsymbol{s}}^t = \frac{1}{J-a} \sum_{i=1}^{J} m_i \boldsymbol{h}_i^t, \tag{6.12}$$

where a is the number of masks, m_i is the binary mask value (*i.e.*, 0 for masking and 1 for unmasking) applied on the i^{th} node representation $\boldsymbol{h}_i^t$, and we have $\sum_{i=1}^{J} m_i = J - a$. With both spatial positional information (Eqs. (6.1), (6.2)) and relational features (Eqs. (6.3), (6.4)) integrated into each node, the masked graph representation $\hat{\boldsymbol{s}}^t \in \mathbb{R}^d$ in Eq. (6.12) retains the context of graph structure, which

is then exploited to prompt the skeleton reconstruction by:

$$\hat{\boldsymbol{x}}^t = f_s(\hat{\boldsymbol{s}}^t), \tag{6.13}$$

where $f_s(\cdot)$ is an MLP network with one hidden layer for skeleton reconstruction with the structure prompt, and $\hat{\boldsymbol{x}}^t \in \mathbb{R}^{J\times 3}$ is the predicted skeleton. It is worth noting that Eq. (6.13) not only utilizes the unmasked node representations to reconstruct their corresponding positions, but also exploits them as the context to predict the masked node positions. For conciseness, we denote the predicted i^{th} training skeleton sequence as $\hat{\boldsymbol{X}}_i = \left(\hat{\boldsymbol{x}}^1, \cdots, \hat{\boldsymbol{x}}^f\right) \in \mathbb{R}^{f\times J\times 3}$.

Graph Trajectory Prompted Reconstruction. To encourage the model to capture more unique temporal patterns from skeleton graphs, we propose to reconstruct graph trajectories based on their partial dynamics. In particular, we randomly mask the trajectory of each node with:

$$\boldsymbol{T}_i = \frac{1}{f-b}\sum_{t=1}^{f} u^t \boldsymbol{h}_i^t, \tag{6.14}$$

where $\boldsymbol{T}_i \in \mathbb{R}^d$ is the *masked* representation of i-node trajectory (*i.e.*, $\boldsymbol{h}_i^1, \cdots, \boldsymbol{h}_i^f$), b is the number of trajectory masks, u^t is the binary mask value applied to the t^{th} position in the i-node trajectory (*i.e.*, i^{th} node representation of the t^{th} skeleton graph), and we have $\sum_{t=1}^{f} u^t = f - b$. The masked trajectory representation $\boldsymbol{T}_i$ preserves partial graph dynamics of different body-joint nodes, which are then used to prompt the skeleton reconstruction by:

$$\overline{\boldsymbol{x}}_i = f_t(\boldsymbol{T}_i). \tag{6.15}$$

In Eq. (6.15), $\overline{\boldsymbol{x}}_i \in \mathbb{R}^{f\times 3}$ denotes the temporal trajectory of the i^{th} body joints predicted by an MLP network $f_t(\cdot)$ with one hidden layer. Based on the unmasked node representations over the temporal dimension, this reconstruction facilitates capturing key temporal dynamics and semantics (*e.g.*, continuous patterns) to infer the node positions missed on the trajectory. We represent the predicted i^{th} training sequence as $\overline{\boldsymbol{X}}_i \in \mathbb{R}^{f\times J\times 3}$ by transposing the original predicted position matrix $(\overline{\boldsymbol{x}}_1, \cdots, \overline{\boldsymbol{x}}_J) \in \mathbb{R}^{J\times f\times 3}$.

We propose the STPR objective to combine both graph structure and trajectory prompted reconstruction as follows:

$$\mathcal{L}_{\text{STPR}} = \beta \mathcal{L}^{st}_{\text{STPR}} + (1-\beta)\mathcal{L}^{tr}_{\text{STPR}}, \tag{6.16}$$

where

$$\mathcal{L}^{st}_{\text{STPR}} = \frac{1}{N_1}\sum_{i=1}^{N_1} \left\| \hat{\boldsymbol{X}}_i - \boldsymbol{X}_i^T \right\|_1, \tag{6.17}$$

$$\mathcal{L}^{tr}_{\text{STPR}} = \frac{1}{N_1}\sum_{i=1}^{N_1} \left\| \overline{\boldsymbol{X}}_i - \boldsymbol{X}_i^T \right\|_1. \tag{6.18}$$

In Eq. (6.16), β is the weight coefficient to fuse structure-prompted ($\mathcal{L}^{st}_{\text{STPR}}$) and trajectory-prompted reconstruction ($\mathcal{L}^{tr}_{\text{STPR}}$). $\boldsymbol{X}_i^T \in \mathbb{R}^{f\times J\times 3}$ denotes the ground-truth positions of i^{th} training skeleton sequence, and $\|\cdot\|_1$ represents the ℓ_1 norm. We employ ℓ_1 reconstruction loss following [8, 114], as it can alleviate gradient explosion of large losses while providing sufficient gradients for the positions with small losses to facilitate more precise reconstruction.

6.2.5 Objective Function of TranSG

We combine the proposed GPC (Sec. 6.2.3) and STPR (Sec. 6.2.4) to perform skeleton representation learning with:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{GPC}} + (1-\lambda)\mathcal{L}_{\text{STPR}}, \tag{6.19}$$

where λ is the weight coefficient to fuse two losses for training. For person re-ID, we leverage the learned SGT to encode skeleton sequences of the probe set Φ_P into sequence-level graph representations, $\{\boldsymbol{S}_i^P\}_{i=1}^{N_2}$ (see Eq. (6.7)), which are matched with the representations, $\{\boldsymbol{S}_i^G\}_{i=1}^{N_3}$, of the same identity in the gallery set Φ_G based on Euclidean distance.

6.3 Experiments

6.3.1 Experimental Setups

6.3.1.1 Datasets

Our approach is evaluated on four skeleton-based person re-ID benchmark datasets: *IAS-Lab RGBD-ID Dataset (IAS)* [120], *KS20 VisLab Multi-View Kinect Skeleton Dataset (KS20)* [119], *BIWI RGBD-ID Dataset (BIWI)* [1], *Kinect Gait Biometry Dataset (KGBD)* [24], which contain 11, 20, 50, and 164 different persons. We also verify the generality of TranSG on the RGB-estimated skeleton data from a large-scale multi-view gait dataset *CASIA-B* [121] with 124 individuals under three conditions (Normal (N), Bags (B), Clothes (C)). We adopt the commonly-used probe and gallery settings following [5, 13] for a fair comparison. Detailed description and examples of these datasets are provided in Sec. 4.3.1.1.

6.3.1.2 Probe and Gallery Settings

We follow the commonly-used settings of probe and gallery in the literature [5]: For the BIWI and IAS datasets, as different testing sets are non-overlapped and contain all pedestrians under different scenes, we evaluate our approach on each testing set by setting it as the probe while the other one is adopted as the gallery. The KGBD dataset contains different skeleton videos (*i.e.*, long skeleton sequences) of each pedestrian with varying numbers of walking rounds. Since no training/testing splits are given, we randomly choose one skeleton video of each person to split skeleton sequences and construct the probe set, and equally divide the remaining videos to build the training set and gallery set. The KS20 dataset collects skeleton data of pedestrians from five different viewpoints, including 0°, 30°, 90°, 130°, and 180°. We employ the setting of Random View Evaluation (RVE): One sequence is randomly selected from each viewpoint as the probe sequence and the remaining skeleton sequences are equally divided into gallery and training sequences. We follow the person re-ID protocols in [13] to evaluate the proposed skeleton-based approach on CASIA-B (detailed in Sec. 6.3.1.3).

6.3.1.3 Evaluation Settings of CASIA-B

In general, 3D skeleton data in existing skeleton-based person re-ID benchmarks are collected with Kinect [22]. To evaluate the effectiveness of our approach when 3D skeleton data are directly estimated from RGB videos rather than depth sensors such as Kinect, we use a large-scale RGB video based dataset, *CASIA-B* [121], which contains walking sequences of 124 individuals under 11 different views and 3 conditions—pedestrians wearing a bag ("Bags"), wearing a coat ("Clothes"), and without any coat or bag ("Normal"). We follow the evaluation setup in [13], which is frequently used in the literature: First, we randomly choose half of the individuals for training and use the rest for testing. Then, to evaluate our approach under *single-condition* and *cross-condition* settings, we divide the testing sequences by the three conditions ("Bags", "Clothes", "Normal") to construct gallery and probe sets. Specifically, for the *single-condition* setting, both gallery and probe sets use the testing sequences with the same condition (*i.e.*, gallery and probe sets are the same), and we match each sequence of the probe set with the most similar sequence from the gallery set that *excludes* the original sequence. In the *cross-condition* setting, we adopt the testing sequences under bags ("Bags") or clothes condition ("Clothes") as the probe set, and use the testing sequences under normal condition ("Normal") as the gallery set.

Following [25], we exploit pre-trained pose estimation models [122, 123] to extract 3D skeletons from RGB videos of CASIA-B. We first extract eighteen 2D joints from each person in videos using the *OpenPose* model [123]. Then, we follow the same configuration of estimation in [25] and average the positions of "Nose", "Reye", "LEye", "Rear" and "Lear" as the position of "Head" to construct fourteen 2D joints, which are fed into the pose estimation method [122] to estimate corresponding 3D body joints. Thus, the number of body-joint nodes J is 14 for CASIA-B as shown in Fig. 6.3, and all joints in each skeleton are normalized by subtracting the neck joint.

6.3.1.4 Implementation Details

Dataset Preprocessing Setups. To avoid ineffective skeleton recording, we discard the first and last 10 skeleton frames of each original skeleton sequence. For KS20, KGBD, BIWI, and IAS datasets, all skeleton sequences are normalized by

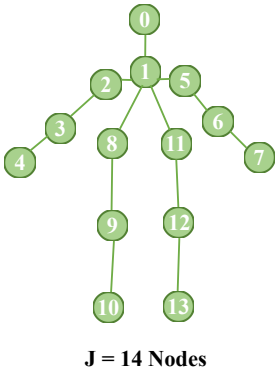


FIGURE 6.3: Node indices for graph representations of the estimated skeletons from CASIA-B dataset. Note: All 3D skeletons are estimated from RGB videos of CASIA-B with [123] and [122] (see Sec. 6.3.1.3).

subtracting the spine joint position from each joint of the same skeleton so that the skeleton is translation invariant [124]. Then, we spilt all normalized skeleton sequences in the training sets into multiple shorter skeleton sequences (*i.e.*, $\boldsymbol{X}$) with length f by a step of $\frac{f}{2}$, which aims to obtain as many 3D skeleton sequences as possible to train our approach. We split all skeleton sequences in the gallery and probe sets into shorter and non-overlapping sequences with length f. Unless explicitly specified, the skeleton sequence $\boldsymbol{X}$ in our work refers to those split and normalized sequences used in learning, rather than those original skeleton sequences provided by datasets. We follow the data augmentation strategy used in [4, 7] to sample more sequences for different identities in the training set, and train our approach with randomly shuffled skeleton sequences of the training set. The details of all datasets are shown in Table 4.1.

Model Parameter Setups. The numbers of body joints are $J = 20$ (IAS, BIWI, KGBD) and $J = 25$ (KS20) in the original datasets. We construct corresponding skeleton graphs with the same number of body-joint nodes in the original skeletons. To verify the generality of our approach when applied to different-scale skeleton graphs, we follow [8] to construct another two scales, namely part-scale (10 nodes) and body-scale (5 nodes), by merging joints within different body partitions. The original skeleton graphs, part-scale graphs, and body-scale graphs as shown in Fig. 6.4 and 6.5. The skeleton sequence length f on four skeleton-based datasets (IAS, KS20, BIWI, KGBD) is set to 6 following [5] for a fair comparison with existing methods. As to CASIA-B, it is a large-scale dataset with roughly estimated skeleton data from RGB frames, which is intrinsically different from the previous datasets. We adopt a longer sequence length $f = 40$. The embedding size of each

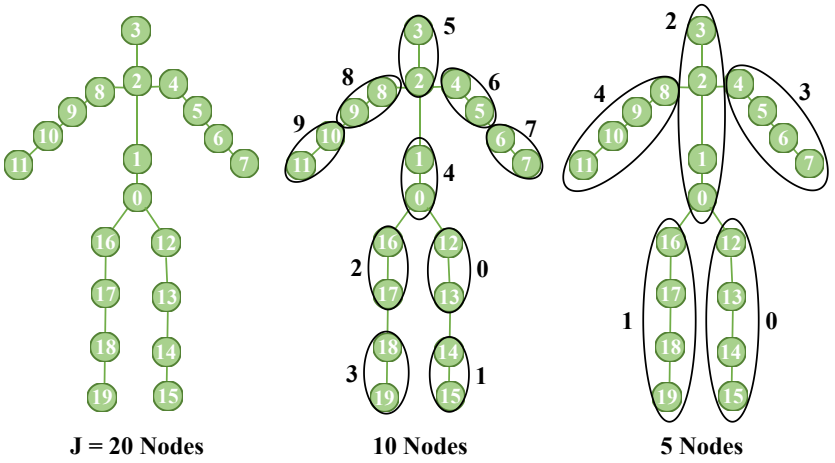


FIGURE 6.4: Node indices for joint-scale (20 nodes), part-level (10 nodes), and body-scale (5 nodes) graphs representations of skeletons from IAS, BIWI and KGBD datasets. Our approach *only* requires joint-scale graphs for training, while we evaluate its performance on different-scale graphs following [8] in our work.

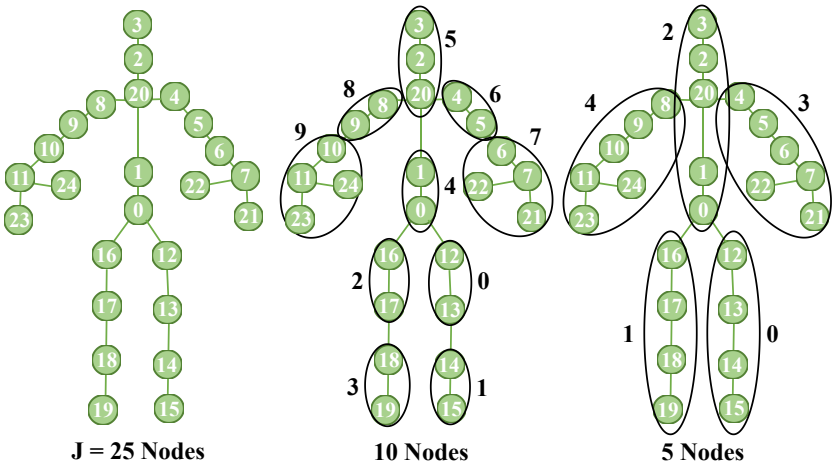


FIGURE 6.5: Node indices for joint-scale (25 nodes), part-level (10 nodes), and body-scale (5 nodes) graphs representations of skeletons from KS20 dataset. Our approach *only* requires joint-scale graphs for training, while we evaluate its performance on different-scale graphs following [8] in our work.

node representation is $d = 128$ for all datasets. We empirically set $K = 10$ for the positional encoding, and employ 2 SGT layers with $H = 8$ attention heads and $d_{\mathrm{k}} = 16$ for each layer, as these settings achieve the best average performance on different datasets. We follow [145, 148] to randomly flip the sign of the eigenvectors during training to improve the model stability on small datasets (IAS, KS20, BIWI). For part-scale (10 nodes) and body-scale (5 nodes) skeleton graphs compared with SM-SGE, we correspondingly set $K = 9$ and $K = 4$ for the positional encoding. For main experiments, we equally fuse each component in our approach

with $\alpha = 0.5, \beta = 0.5, \lambda = 0.5$. For the experiments with RGB-estmiated skeletons, we empirically set $\alpha = 1.0$ as it can achieve better performance. We empirically set $\tau_1 = 0.07$ and $\tau_2 = 14$ for contrastive learning, while using $a = 10$ and $b = 2$ random masks for STPR. An Adam optimizer with the learning rate of of 3.5×10^{-4} is used for the model optimization, and we set batch size to 256 for all datasets. To apply our approach to unsupervised skeleton representation learning without using ground-truth labels, we follow [61] to perform DBSCAN clustering [6] of graph representations, and leverage their pseudo classes to generate graph prototypes for contrastive learning. To avoid over-fitting and achieve better generalization performance, we adopt Early Stopping [126] with a patience of 120 epochs (*i.e.*, stop the training of model after no improvement in 120 continuous epochs). The experiments are repeated for multiple time with random model parameter initialization for training, and we report the average performance for a fair comparison with existing methods. Interested readers can access our source codes[2] for more details.

Method Comparison Setups. For all methods compared in our experiments, we select optimal model parameters for training, and use their pre-defined skeleton descriptors or pre-trained skeleton representations for person re-ID. It is worth noting that our re-implementations of some existing models get performance with slight variations, and the results are basically the same as the original papers under different random model initializations. For a fair comparison, we follow [4, 5] to report the average performance of all methods. Note that our approach does not use any post-processing technique, *e.g.*, re-ranking [127] or multi-query fusion [49] in the training or testing stage. To perform person re-ID, we exploit the approach to encode each original skeleton sequence of the probe set Φ_P into corresponding sequence-level graph representations, $\{\boldsymbol{S}_i^P\}_{i=1}^{N_2}$, and match it with representations, $\{\boldsymbol{S}_i^G\}_{i=1}^{N_3}$, of the same identity in the gallery set Φ_G using Euclidean distance. In the ablation study, we use the concatenation of raw skeleton sequences (*i.e.*, normalized 3D coordinates of body joints) as the baseline. For the configuration of naïve prototype contrastive learning (PC), we adopt the same setting in [5]: We leverage DBSCAN [6] to cluster original skeleton sequences in an unsupervised manner, and directly use the feature centroid in each cluster as the skeleton prototype for contrastive learning.

[2]Our codes are publicly available at https://github.com/Kali-Hac/TranSG.

TABLE 6.1: Performance comparison with state-of-the-art skeleton-based methods on different datasets. ♠ denotes skeleton graph based methods, † indicates using hand-crafted descriptors, and ‡ refers to sequence representation learning models. **Bold** denotes the best results.

Methods	BIWI-S				BIWI-W				KS20				IAS-A				IAS-B				KGBD			
	mAP	R_1	R_5	R_{10}	mAP	R_1	R_5	R_{10}	mAP	R_1	R_5	R_{10}	mAP	R_1	R_5	R_{10}	mAP	R_1	R_5	R_{10}	mAP	R_1	R_5	R_{10}
D_{13}† [1]	13.1	28.3	53.1	65.9	17.2	14.2	20.6	23.7	18.9	39.4	71.7	81.7	24.5	40.0	58.7	67.6	23.7	43.7	68.6	76.7	1.9	17.0	34.4	44.2
D_{16}† [2]	16.7	32.6	55.7	68.3	18.8	17.0	25.3	29.6	24.0	51.7	77.1	86.9	25.2	42.7	62.9	70.7	24.5	44.5	69.1	80.2	4.0	31.2	50.9	59.8
PoseGait†[25]	9.9	14.0	40.7	56.7	11.1	8.8	23.0	31.2	23.5	49.4	80.9	**90.2**	17.5	28.4	55.7	69.2	20.8	28.9	51.6	62.9	13.9	50.6	67.0	72.6
AGE‡[3]	8.9	25.1	43.1	61.6	12.6	11.7	21.4	27.3	8.9	43.2	70.1	80.0	13.4	31.1	54.8	67.4	12.8	31.1	52.3	64.2	0.9	2.9	5.6	7.5
SGELA‡[4]	15.1	25.8	51.8	64.4	19.0	11.7	14.0	14.7	21.2	45.0	65.0	75.1	13.2	16.7	30.2	44.0	14.0	22.2	40.8	50.2	4.5	38.1	53.5	60.0
MG-SCR♠[7]	7.6	20.1	46.9	64.1	11.9	10.8	20.3	29.4	10.4	46.3	75.4	84.0	14.1	36.4	59.6	69.5	12.9	32.4	56.5	69.4	6.9	44.0	58.7	64.6
SM-SGE♠[8]	10.1	31.3	56.3	69.1	15.2	13.2	25.8	33.5	9.5	45.9	71.9	81.2	13.6	34.0	60.5	71.6	13.3	38.9	64.1	75.8	4.4	38.2	54.2	60.7
SPC-MGR♠[61]	16.0	34.1	57.3	69.8	19.4	18.9	31.5	40.5	21.7	59.0	79.0	86.2	24.2	41.9	66.3	75.6	24.1	43.3	68.4	79.4	6.9	40.8	57.5	65.0
SimMC‡[5]	12.3	41.7	66.6	76.8	19.9	24.5	36.7	44.5	22.3	66.4	80.7	87.0	18.7	44.8	65.3	72.9	22.9	46.3	68.1	77.0	11.7	54.9	66.2	70.6
Hi-MPC‡[133]	17.4	47.5	70.3	78.6	22.6	27.3	40.3	48.8	22.0	69.6	83.5	87.1	23.2	45.6	67.3	75.4	25.3	48.2	70.2	77.8	10.2	56.9	70.2	75.1
TranSG♠ (Ours)	**30.1**	**68.7**	**86.5**	**91.8**	**26.9**	**32.7**	**44.9**	**52.2**	**46.2**	**73.6**	**86.3**	**90.2**	**32.8**	**49.2**	**68.5**	**76.2**	**39.4**	**59.1**	**77.0**	**87.0**	**20.2**	**59.0**	**73.1**	**78.2**

6.3.1.5 Evaluation Metrics

We compute the Cumulative Matching Characteristics (CMC) curve and adopt Rank-1 accuracy (R_1), Rank-5 accuracy (R_5), and Rank-10 accuracy (R_{10}) as performance metrics. R_1, R_5, and R_{10} are computed as the ratios of probe sequences correctly identified by matching with only the top 1, within top 1 to 5, and within top 1 to 10 most similar candidate gallery sequences, respectively. Mean Average Precision (mAP) [49] is also used to quantitatively evaluate the overall performance of our approach.

6.3.2 Comparison with State-of-the-Art Methods

We compare our approach with state-of-the-art graph-based methods, hand-crafted methods, and sequence learning methods on BIWI, KS20, IAS, and KGBD in Table 6.1.

6.3.2.1 Comparison with Graph-based Methods

As shown in Table 6.1, the proposed TranSG significantly outperforms two state-of-the-art graph-based methods, MG-SCR [7] and SM-SGE [8], by 11.7-36.7% for mAP and 12.8-48.6% for Rank-1 accuracy on different datasets. As these methods rely on multi-stage relation modeling and sequence-level context learning, the results demonstrate that the proposed full-relation learning model (SGT) with fine-grained (*i.e.*, graph and node level) semantics learning (STPR) is able to learn

TABLE 6.2: Ablation study with different configurations: Naïve prototype contrastive learning (PC), skeleton graph transformer (SGT) with direct supervised learning (DS) or graph prototype contrastive learning (GPC), and structure-trajectory prompted reconstruction (STPR). "+" indicates employing the component.

Configurations	BIWI-S		BIWI-W		KS20		KGBD	
	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1
Baseline	9.3	24.8	14.1	10.9	9.5	17.0	6.4	34.5
PC	11.3	38.1	18.3	21.2	20.5	64.8	11.0	53.0
SGT + DS	19.0	42.4	21.1	21.7	27.6	60.0	11.1	51.5
SGT + GPC	26.7	66.6	25.5	31.2	42.5	71.3	18.1	57.0
SGT + GPC + STPR	30.1	68.7	26.9	32.7	46.2	73.6	20.2	59.0

more unique skeleton features for person re-ID. Compared with the latest SPC-MGR model [61] that utilizes sequence-level graph representations for contrastive learning, our model achieves superior performance with a large margin of 7.5% to 24.5% for mAP and 7.3% to 34.6% for Rank-1 accuracy on all datasets. This suggests the higher effectiveness of our approach combining both sequence-level and skeleton-level graph prototype contrastive learning. We will also show the generality of our model under different-scale skeleton graph modeling in Sec. 6.4.

6.3.2.2 Comparison with Hand-crafted and Sequence Learning Methods

In contrast to the methods using hand-crafted anthropometric descriptors (D_{13} [1], D_{16} [2]) or 3D pose features (PoseGait [25]), our approach consistently achieves higher performance by up to 27.3% for mAP and 54.7% for Rank-1 accuracy on all datasets. TranSG also achieves a remarkable improvement over existing sequence representation learning models (AGE [3], SGELA [4], SimMC [5]) in terms of mAP (7.0-37.3%), Rank-1 (4.1-56.1%), Rank-5 (3.2-67.5%), and Rank-10 accuracy (3.2-70.7%). This demonstrates the superiority of our graph-based model, as it can fully capture body relations and discriminative patterns from skeleton graphs for the person re-ID task.

TABLE 6.3: Performance comparison with *appearance* or *skeleton* based methods on CASIA-B. "B-N" denotes matching the "Bags (B)" probe set with the"Normal (N)" gallery set. "—" indicates no published result. Full results are provided in Appendix II.

	Probe-Gallery	N-N		B-B		C-C		C-N		B-N	
	Methods	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1
Appearance	LMNN [9]	—	3.9	—	18.3	—	17.4	—	11.6	—	23.1
	ITML [10]	—	7.5	—	19.5	—	20.1	—	10.3	—	21.8
	ELF [11]	—	12.3	—	5.8	—	19.9	—	5.6	—	17.1
	SDALF [12]	—	4.9	—	10.2	—	16.7	—	11.6	—	22.9
	MLR (Scores) [13]	—	13.6	—	13.6	—	13.5	—	9.7	—	14.7
	MLR (Features) [13]	—	16.3	—	18.9	—	25.4	—	20.3	—	31.8
Skeleton	AGE [3]	3.5	20.8	9.8	37.1	9.6	35.5	3.0	14.6	3.9	32.4
	SM-SGE [8]	6.6	50.2	9.3	26.6	9.7	27.2	3.0	10.6	3.5	16.6
	SPC-MGR [61]	9.1	71.2	11.4	44.3	11.8	48.3	4.3	22.4	4.6	28.9
	SGELA [4]	9.8	71.8	16.5	48.1	7.1	51.2	4.7	15.9	6.7	36.4
	SimMC [5]	10.8	**84.8**	16.5	**69.1**	**15.7**	**68.0**	5.4	**25.6**	7.1	42.0
	TranSG (Ours)	**13.1**	78.5	**17.9**	67.1	**15.7**	65.6	**6.7**	23.0	**8.6**	**44.1**

6.4 Further Analysis

6.4.1 Ablation Study

We conduct ablation study to verify the effectiveness of each component in our approach. The skeleton sequences of concatenated joints are adopted as the baseline, and we include the naïve prototype contrastive learning (PC) using original sequences [5] for comparison. As shown in Table 6.2, compared with using raw skeleton sequences or PC without graph modeling, applying SGT achieves significantly better performance in most cases, regardless of using GPC or not. This demonstrates the effectiveness of the skeleton graph learning with SGT, as it can fully capture relations within skeletons to learn unique body structure and motion features for person re-ID. The SGT employing GPC achieves superior results than "SGT + DS" that uses direct supervised learning (*i.e.*, cross-entropy loss) with an improvement of 4.4-14.9% for mAP and 5.5-24.2% for Rank-1 accuracy, which verifies the key role of the graph prototype contrastive learning in capturing more representative discriminative graph features from different identities. Finally, adding STPR consistently boosts model performance by 1.4-3.7% for mAP on all dataset, which further demonstrates its effectiveness on capturing more valuable graph semantics and discriminative patterns for person re-ID.

TABLE 6.4: Performance of our approach on different-scale skeleton graphs. J, 10 or 5 nodes correspond to the joint-scale, part-scale or body-scale skeletal structures used in SM-SGE [8].

Nodes	Methods	BIWI-S		BIWI-W		KS20		KGBD	
		mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1
J	SM-SGE [8]	10.0	33.0	14.9	12.9	10.2	44.7	4.3	40.2
	TranSG (Ours)	**30.1**	**68.7**	**26.9**	**32.7**	**46.2**	**73.6**	**20.2**	**59.0**
10	SM-SGE	11.1	32.8	16.5	14.5	9.8	43.2	4.1	33.0
	TranSG (Ours)	**15.1**	**37.9**	**18.7**	**20.0**	**17.0**	**48.1**	**4.9**	**36.9**
5	SM-SGE	10.0	27.5	13.8	12.6	9.3	37.3	4.4	31.5
	TranSG (Ours)	**13.5**	**36.6**	**14.1**	**17.0**	**13.2**	**40.8**	**4.6**	**34.6**

6.4.2 Evaluation on RGB-estimated Skeletons

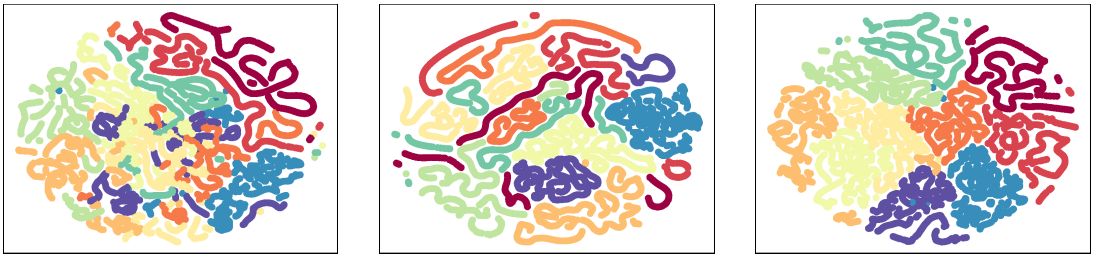
To verify the generality of our skeleton-based model on RGB-estimated skeletons, we utilize pre-trained pose estimation models to extract skeleton data from RGB videos of CASIA-B [25], and evaluate the performance of TranSG under different conditions. As presented in Table 6.3, our approach not only outperforms many existing state-of-the-art skeleton-based models with a prominent improvement in most conditions, but also achieves superior performance to representative classical appearance-based methods that utilize RGB-based features (*e.g.*, textures, silhouettes) or/and visual metric learning [9–13]. This demonstrates the stronger ability of TranSG on capturing more discriminative features from the estimated skeletons, and also suggests its great potential to be applied to more general RGB-based scenarios.

6.4.3 Evaluation on Skeleton Graphs with Varying Scales

To validate the effectiveness of our approach under different graph modeling, we follow [8] to construct different-scale graphs for model learning. As shown in Table 6.4, TranSG significantly outperforms the state-of-the-art framework SM-SGE [8] when utilizing the original skeleton graphs (corresponding to J joints) or higher level graphs with less nodes. This demonstrates that our model is compatible with different-level skeletal structures and can learn more effective features even under different graph modeling.

TABLE 6.5: Performance of our approach using only unlabeled skeletons. We use the same clustering setting in [61] for comparison.

Methods	BIWI-S		BIWI-W		KS20		KGBD	
	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1
SPC-MGR [61]	**16.0**	34.1	19.4	18.9	21.7	59.0	6.9	40.8
TranSG (Ours)	14.2	**42.2**	**21.9**	**26.6**	**23.6**	**65.0**	**7.5**	**43.6**



(A) SPC-MGR [61] (B) SimMC [5] (C) TranSG (Ours)

FIGURE 6.6: t-SNE visualization of representations learned by SPC-MGR (a), SimMC (b), and TranSG (c) for first ten classes in IAS. Different colors indicates representations of different classes.

6.4.4 Evaluation in Unsupervised Scenarios

To apply our approach in an unsupervised manner without using ground-truth labels, we follow [61] to perform DBSCAN clustering [6] of graph representations, and leverage their pseudo classes to generate graph prototypes for contrastive learning. With only *unlabeled* skeletons as inputs, TranSG can still achieve superior performance than the latest graph-based method SPC-MGR [61] in most cases, as shown in Table 6.5. This further demonstrates the generality of our approach, which could be promisingly transferred to more related tasks such as unsupervised open-set person re-ID.

6.4.5 Feature Visualization

As shown in Fig. 6.6, we conduct a t-SNE visualization [128] of skeleton representations. The result shows that our learned skeleton representations possess more discriminative inter-class separation than other methods SPC-MGR [61] and SimMC [5], which indicates that TranSG may capture richer class-related semantics.

TABLE 6.6: The number of network parameters (million (M)) and computational complexity (giga floating-point operations (GFLOPs)) of deep learning based methods. ♠ denotes skeleton graph based methods, † indicates using hand-crafted descriptors, and ‡ refers to sequence representation learning models. Note: Both numbers of parameters and GFLOPs in the training of neural networks are counted by the Tensorflow platform [136]. Extra matrix computation is required for the clustering in SimMC and SPC-MGR (see Sec. 5.4.6).

Methods	# Params	GFLOPs
PoseGait†[25]	8.93M	121.60
MG-SCR♠[7]	0.35M	6.60
AGE‡[3]	7.15M	37.37
SGELA‡[4]	8.47M	7.47
SM-SGE♠[8]	5.58M	22.61
SPC-MGR♠[61]	0.01M	0.12
SimMC‡[5]	0.15M	0.99
TranSG♠ (Ours)	0.40M	20.22

6.4.6 Model Efficiency

We report the model efficiency in terms of model size, *i.e.*, number of network parameters, and computational complexity for existing deep learning based methods. For the model that possesses varying sizes and complexities on different datasets due to the changes of input data, we report the largest case. As shown in Table 6.6, the proposed approach possesses smaller model size than many existing skeleton-based person re-ID methods (PoseGait [25], AGE [3], SGELA [4], SM-SGE [8]). The number of GFLOPs in Table 6.6 refers to computational complexity in the training of neural networks, which is the whole[3]/main computational complexity for deep learning methods. It should be noted that the unsupervised prototype contrastive learning (SimMC [5], SPC-MGR [61]) requires extra matrix computation (*e.g.*, vector similarity query) for the clustering process , which is usually time-consuming and computationally expensive as it may require both CPU and GPU (*e.g.*, using Faiss library [137]). In contrast, our approach exploits ground-truth identities to generate graph prototypes, which can not only improve the prototype reliability but also achieve significantly faster training without requiring clustering.

[3]For representation learning methods without other learning processes (*e.g.*, clustering), the whole computational complexity of the model can be equivalent to the computational complexity of the used neural networks.

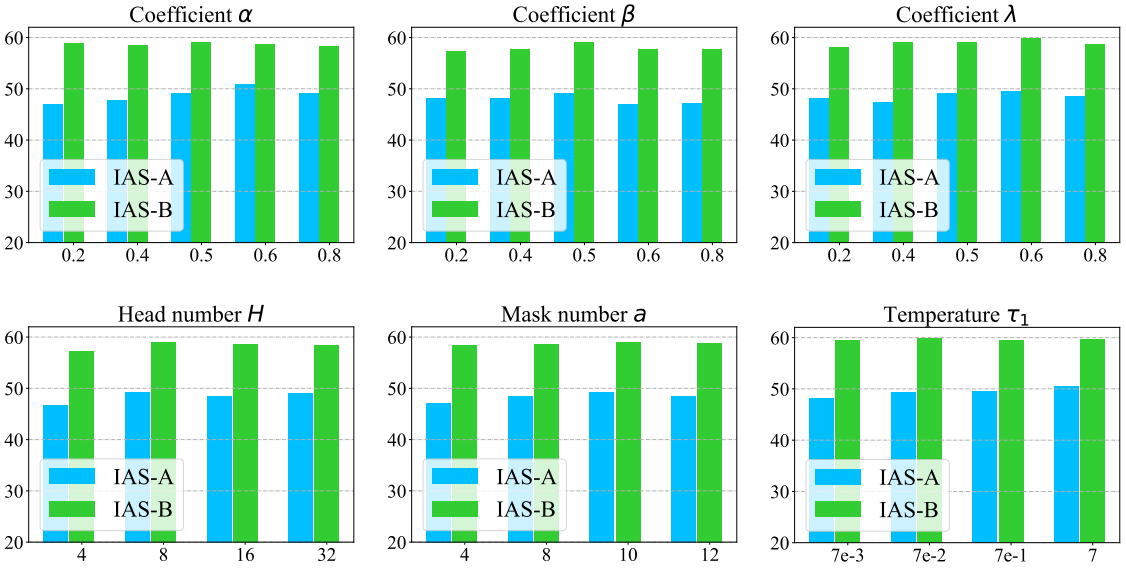


FIGURE 6.7: Rank-1 accuracy of our approach on different probe sets (IAS-A and IAS-B) when setting different hyper-parameters.

TABLE 6.7: Performance of our approach on different datasets when setting different weight coefficients α to fuse sequence-level ($\mathcal{L}_{\text{GPC}}^{seq}$) and skeleton-level graph prototype contrastive learning ($\mathcal{L}_{\text{GPC}}^{ske}$) in the proposed GPC.

α	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
0.0	71.0	33.5	55.5	14.2	47.3	29.7	57.3	36.3	27.3	25.5	62.7	21.3
0.2	71.5	40.0	56.3	17.9	46.9	31.2	58.9	38.6	31.2	27.5	67.2	29.7
0.4	72.2	45.7	57.5	18.6	47.7	30.9	58.4	38.0	31.7	26.0	68.7	29.9
0.5	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
0.6	73.1	46.3	58.0	19.0	50.9	33.6	58.7	38.8	31.0	26.5	66.8	26.7
0.8	71.7	44.0	58.8	19.8	49.2	31.5	58.2	37.8	32.0	26.0	64.7	26.7
1.0	70.1	41.8	56.9	18.0	47.3	32.6	54.3	37.8	30.3	25.2	64.5	26.1

6.4.7 Analysis of Hyperparameters

Overview. As presented in Fig. 6.7, we show effects of different hyper-parameters on our approach. An appropriate fusion of different components can encourage better model performance, while setting α, β, and λ to 0.5 achieves slightly better results. Adding too many FR heads or SGT layers could slightly reduce the performance, as it might expand the model scale and learn more redundant information. Our model is not sensitive to the variation of some parameters such as temperature τ_1, while setting a moderate value for mask numbers benefits model performance. We provide more detailed results and analyses in this section.

TABLE 6.8: Performance of our approach on different datasets when setting different weight coefficients β to combine graph trajectory-prompted ($\mathcal{L}^{tr}_{\text{STPR}}$) and structure-prompted reconstruction ($\mathcal{L}^{st}_{\text{STPR}}$) in the proposed STPR.

β	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
0.0	73.0	45.4	58.6	19.6	48.4	32.1	58.2	39.2	31.1	25.9	67.8	28.2
0.2	72.6	44.5	56.4	17.8	48.2	32.3	57.3	39.1	31.8	27.0	68.0	29.6
0.4	72.7	44.6	57.1	21.1	48.1	32.4	57.6	40.1	31.5	27.0	69.3	30.9
0.5	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
0.6	72.9	44.7	56.1	18.4	46.9	33.2	57.7	38.7	33.1	27.9	68.6	30.4
0.8	72.3	47.1	58.8	20.4	47.1	32.9	57.6	39.9	31.0	26.3	68.7	30.4
1.0	71.9	44.8	58.1	18.9	48.7	32.4	57.2	38.1	31.9	26.2	67.6	27.0

TABLE 6.9: Performance of our approach on different datasets when setting different weight coefficients λ to fuse the graph prototype contrastive learning ($\mathcal{L}_{\text{GPC}}$) and structure-trajectory prompted reconstruction ($\mathcal{L}_{\text{STPR}}$) for model training.

λ	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
0.0	44.9	16.8	29.8	4.3	35.3	23.2	36.4	25.9	18.8	16.0	35.9	15.3
0.2	72.9	45.9	56.4	18.4	48.2	34.0	58.1	39.4	33.5	27.0	67.4	30.3
0.4	72.4	45.1	58.7	21.0	47.4	32.3	59.1	39.9	31.4	26.3	69.1	30.6
0.5	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
0.6	71.6	44.2	58.2	19.6	49.5	33.1	59.8	40.0	32.0	24.6	66.4	30.8
0.8	71.8	41.7	58.1	19.9	48.5	32.7	58.7	39.3	32.4	28.2	65.5	26.5
1.0	71.3	42.5	57.0	18.1	48.0	33.0	56.1	40.2	31.2	26.2	66.6	26.7

TABLE 6.10: Performance of our approach on different datasets when setting different numbers of Skeleton Graph Transformer (SGT) layers.

Layer	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
1	69.8	39.0	58.6	21.0	46.8	31.8	58.5	40.3	28.8	25.6	65.8	27.1
2	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	67.8	30.1
3	72.7	41.0	53.3	13.9	49.7	33.7	57.7	40.1	33.2	27.8	66.2	30.0
4	69.0	36.2	52.8	11.8	46.6	29.8	57.0	38.1	31.1	26.1	63.3	24.3

TABLE 6.11: Performance of our approach on different datasets when setting different numbers of attention heads per SGT layer.

H	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
4	73.2	46.2	56.9	18.4	46.7	33.9	57.2	38.0	31.2	26.5	67.5	29.9
8	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
16	72.7	46.4	56.6	17.5	48.5	33.0	58.6	39.5	32.7	26.7	68.5	30.2
32	72.9	44.7	52.6	15.4	49.1	32.0	58.5	39.9	32.5	27.0	66.3	29.5

6.4.7.1 Effects of Different Weight Coefficient α, β, and λ

As shown in Table 6.9, our approach can achieve the best performance in average on different datasets when equally (*i.e.*, $\lambda = 0.5$) fusing the proposed graph prototype contrastive learning (GPC) and structure-trajectory prompted reconstruction (STPR). This is also consistent with our analysis in Sec. 6.4.11 that GPC and

TABLE 6.12: Performance of our approach on different datasets when setting different temperature τ_1.

τ_1	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
0.01	73.1	42.2	59.2	18.1	48.1	32.0	59.4	40.7	31.9	26.6	68.0	27.2
0.1	73.4	45.2	58.0	20.4	49.3	32.6	59.8	41.8	31.5	27.1	69.3	29.5
1.0	72.7	43.9	57.8	19.6	49.6	35.8	59.1	39.6	32.3	27.0	69.1	32.1
10	73.4	41.1	57.0	18.4	50.5	32.0	59.7	40.3	32.7	26.7	66.7	30.2

TABLE 6.13: Performance of our approach on different datasets when setting different temperature τ_2.

τ_2	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
0.01	70.1	31.0	55.5	13.2	49.5	31.9	55.6	35.1	32.8	26.8	61.1	15.5
0.1	72.3	39.1	56.8	13.1	49.3	29.8	57.8	36.7	29.6	26.6	67.9	27.2
1.0	73.2	43.3	57.0	16.2	48.4	33.3	59.1	37.8	32.8	27.5	68.3	29.6
10	73.4	46.0	58.7	18.5	49.0	32.6	59.0	40.9	33.0	27.2	69.0	31.3

TABLE 6.14: Performance of our approach on different datasets when setting different numbers of random structure masks.

a	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
4	72.2	46.3	58.0	19.8	47.2	31.3	58.4	39.6	32.9	28.8	67.8	30.2
8	73.2	47.0	58.2	19.5	48.5	32.9	58.6	41.8	33.5	27.4	69.3	31.0
10	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
12	72.7	45.7	58.8	20.3	48.4	32.6	58.8	40.6	33.1	27.0	69.0	30.5

TABLE 6.15: Performance of our approach on different datasets when setting different numbers of random trajectory masks.

b	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
1	72.7	45.3	59.1	20.0	48.7	31.9	57.5	38.6	32.5	26.6	69.2	31.8
2	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
3	73.4	47.3	59.3	19.9	48.5	33.2	58.1	39.6	32.6	27.0	68.5	30.4
4	72.9	44.9	58.7	19.8	49.3	32.5	57.5	38.8	33.0	28.4	69.0	31.8

TABLE 6.16: Performance of our approach on different datasets when employing different sequence length f.

f	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
4	70.1	47.5	59.3	22.9	49.1	35.3	56.4	35.1	32.3	27.2	66.8	27.5
6	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
8	75.8	52.6	58.3	18.9	47.9	41.7	61.2	43.7	32.2	33.8	76.6	39.1
10	80.5	49.3	58.5	20.9	49.6	37.2	57.3	50.0	34.8	34.1	63.3	37.4

STPR are compatible and can facilitate each other to learn better skeleton graph representations for person re-ID. Interestingly, only using the reconstruction mechanism (STPR) without GPC ($\lambda = 0.0$) can still learn effective skeleton graph features for person re-ID despite with significantly lower accuracy, which suggests the higher contribution of GPC and the limited ability of STPR on learning discriminative skeleton features. For GPC, we observe that an appropriate fusion

TABLE 6.17: Performance of our approach on different datasets when the SGT uses (w/) positional encoding or without (w/o) positional encoding.

Pos. Enc.	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
w/	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
w/o	74.1	48.3	49.2	13.2	44.8	29.1	53.3	37.6	32.6	27.6	66.1	28.4

TABLE 6.18: Performance of our approach on different datasets when the STPR uses ℓ_1 or ℓ_2 loss.

Loss Type	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
ℓ_2	48.1	15.8	36.0	5.7	46.5	24.1	48.3	35.5	33.0	29.0	61.3	29.5
ℓ_1	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1

($\alpha = 0.4 - 0.6$) of sequence-level ($\mathcal{L}_{\text{GPC}}^{seq}$) and skeleton-level ($\mathcal{L}_{\text{GPC}}^{ske}$) prototype contrastive learning obtains better results than solely using them (*i.e.*, $\alpha = 0.0$ or $\alpha = 1.0$), as shown in Table 6.7. Our model is not sensitive to the changes of β when fusing structure-prompted and trajectory-prompted reconstruction. As shown in Table 6.8, $\beta = 0.5$ achieves slightly better performance in average, while a smaller value of β could benefit the model performance on some datasets such as BIWI-S. As the skeleton data of different domains (*e.g.*, datasets) are collected under different conditions, the context of skeleton structure or trajectory may have different contributions on the reconstruction and skeleton semantics learning, thus β can be further selected to facilitate the model training.

6.4.7.2 Effects of Different Numbers of Attention Heads and SGT Layers

As shown in Tables 6.10 and 6.11, setting 2 SGT layers and 8 attention heads per layer enables our model to obtain the best performance on different datasets. Employing too many numbers of layers (4 layers) slightly reduces the performance as it may largely expand the model scale and learn more redundant information. The results also show that our model trained on the large dataset such as KGBD is more sensitive to the layer variation. Since our approach under different numbers of attention heads achieves similar performance, we empirically select $H = 8$ heads to achieve a better trade-off between computational cost and performance.

6.4.7.3 Effects of Other Parameters

As shown in Tables 6.12, 6.14 and 6.15, our approach is not sensitive to changes of some parameters such as the temperature τ_1 and random mask numbers (a, b). In practice, we select $a = 10$ and $b = 2$ as this setting achieves slightly better performance on different datasets. Although setting different τ_1 value may obtain similar results, we observe that their scales could influence the training stability (*i.e.*, setting too small or too large values induces more evident performance variations). We therefore choose a moderate value for the temperature τ_1. The results in 6.13 show that TranSG achieves higher performance when setting a relatively higher value for the temperature τ_2, which also improves the training stability (*i.e.*, smaller loss fluctuation) of our model on different datasets. In our experiments, the temperatures are empirically set to $\tau_1 = 0.07$ and $\tau_2 = 14$, and they could be further tuned for better performance.

Analysis of ℓ_1 loss for STPR. As shown in Table 6.18, employing ℓ_1 loss for STPR mechanism achieves significantly higher performance than using ℓ_2 loss on different datasets. It is also observed that using ℓ_2 could induce relatively large performance variations in terms of both Rank-1 accuracy and mAP, especially on large datasets such as KS20 and KGBD. These results suggest that solely using the ℓ_2 loss might be difficult to learn precise or stable reconstruction in TranSG, and verify the effectiveness of ℓ_1 norm in facilitating better higher model performance with more stable structure-trajectory prompted reconstruction, which is also consistent with the analysis in previous work[114]. It is worth noting that in practice the deep learning platforms such as Tensorflow have manually defined a pseudo gradient (e.g., gradient is set to 0) for ℓ_1 loss at x=0 to avoid the non-differentiable problem in loss optimization.

6.4.8 Analysis of Multi-Shot Performance

We evaluate the multi-shot performance of our approach with different settings of sequence lengths f (*i.e.*, f-shot person re-ID). Since skeleton sequences contain more pattern features as f increases, our approach is capable of learning more effective skeleton graph representations to achieve larger performance improvement in most cases as shown in 6.16. Nevertheless, it is interesting to note that using shorter sequences sometimes performs better than longer sequences on small

datasets such as IAS-B and BIWI-S, implying that a larger size of available training sequences under smaller f settings could help learn better representations on those datasets. It should be noted that in our work, we evaluate all compared methods under the same sequence length ($f = 6$) following the literature [5, 61].

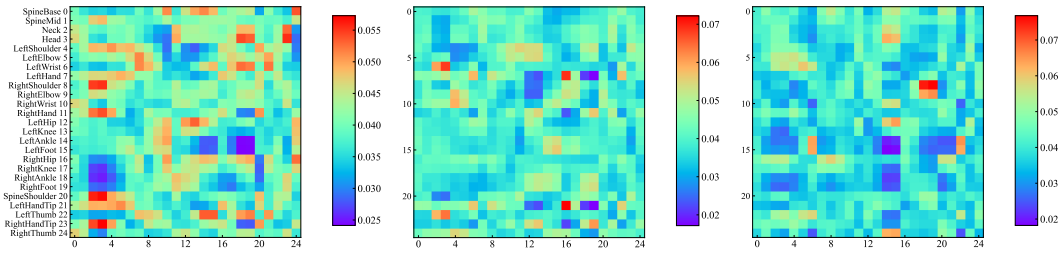
6.4.9 Analysis of Positional Encoding

The positional encoding used in our SGT helps preserve the unique positional information of nodes based on the graph structure *i.e.*, structurally nearby nodes are endowed with similar positional features while the farther nodes possess more different positional features [145, 148]. As shown in Table 6.17, removing the positional encoding can reduce our model performance in most cases, which demonstrates the important role of positional information in the skeleton graph learning of our approach, as it encourages capturing richer structural graph context for relation learning and graph reconstruction. Interestingly, our model achieves similar (slightly lower) performance on the KS20 dataset when removing positional encoding. As each skeleton of KS20 contains more body-joint nodes than that of other datasets, there could be two possible reasons for this result: (1) The positional encoding based on a small number of eigenvectors ($K = 10$) might be insufficient to characterize the unique node information from a large skeleton graph, and we can improve K for better learning. (2) Directly introducing positional information into the learning of larger graphs might not be suitable, so a more effective skeleton-based positional encoding mechanism should be devised.

In practice, we found that setting different K ranging from 6 to 14 achieves similar performance, while using $K = 10$ obtains slightly better average performance on different datasets. The results show that our model is not sensitive to the change of eigenvector number. We will further explore the issue (2) in the future work.

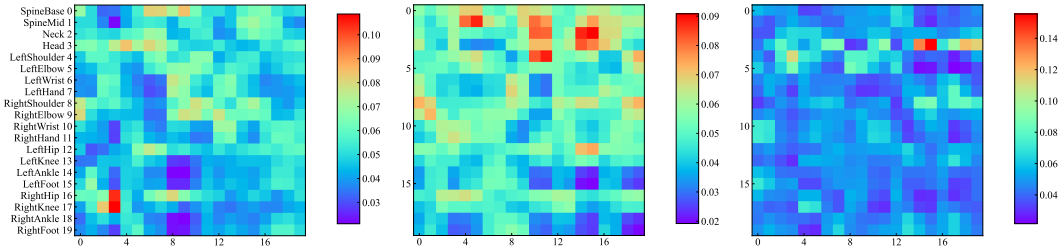
6.4.10 Analysis of Body Relations

As shown in Fig. 6.8-6.11, we visualize three learned full-relation (FR) matrices for the same skeleton sequence in different datasets. Note that there are totally $H = 8$ learned relation matrices in our approach and here we only visualize 3 of them. Since each FR head computes f relation matrices corresponding to f skeleton



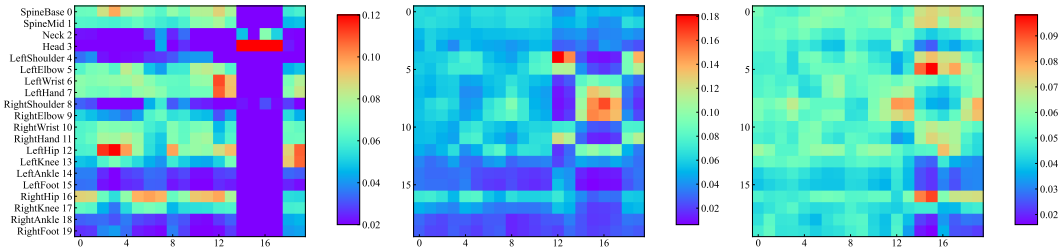
(A) (B) (C)

FIGURE 6.8: (a) Visualization of full-relation (FR) among body-joint nodes for testing skeletons in KS20. (a)-(b) represent the relations learned by the 1^{st}, 4^{th}, and 8^{th} FR heads. Note that the abscissa and ordinate denote indices of nodes (see Sec. 6.3.1.4).



(A) (B) (C)

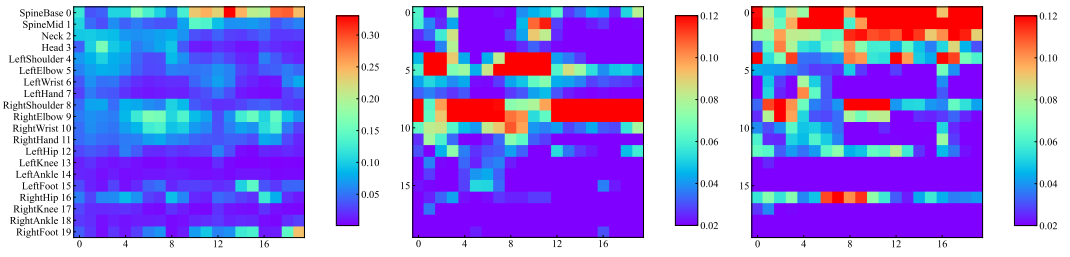
FIGURE 6.9: (a) Visualization of full-relation (FR) among body-joint nodes for testing skeletons in IAS. (a)-(b) represent the relations learned by the 1^{st}, 4^{th}, and 8^{th} FR heads. Note that the abscissa and ordinate denote indices of nodes (see Sec. 6.3.1.4).



(A) (B) (C)

FIGURE 6.10: (a) Visualization of full-relation (FR) among body-joint nodes for testing skeletons in BIWI. (a)-(b) represent the relations learned by the 1^{st}, 4^{th}, and 8^{th} FR heads. Note that the abscissa and ordinate denote indices of nodes (see Sec. 6.3.1.4).

graphs in a sequence, we average them into a matrix to show the mean relations of body-joint nodes. We can observe that FR heads can capture different correlations between different nodes, and they can individually focus on patterns of the same body part correlated with other parts. For example, the 4^{th} FR head trained on BIWI focuses on salient relations between nodes 6-9 and nodes 14-17 (see Fig. 6.10



(A) (B) (C)

FIGURE 6.11: (a) Visualization of full-relation (FR) among body-joint nodes for testing skeletons in KGBD. (a)-(b) represent the relations learned by the 1^{st}, 4^{th}, and 8^{th} FR heads. Note that the abscissa and ordinate denote indices of nodes (see Sec. 6.3.1.4).

(b)), while the 8^{th} head pays more attention to patterns between nodes 6-9 and other body components (*i.e.*, nodes 12-13 and 18-19 (see Fig. 6.10 (c)). These results demonstrate that the multiple FR heads in SGT can capture different body and motion relations of nodes from different representation subspaces to facilitate learning a better skeleton graph representation.

6.4.11 Analysis of Training Process

We visualize the total training loss $\mathcal{L}$ in Fig. 6.12, and the results suggest that our model learning can converge very fast in the first 100 optimization epochs. Meanwhile, the graph prototype contrastive (GPC) loss $\mathcal{L}_{\text{GPC}}$ and structure-trajectory prompted reconstruction (STPR) loss curves $\mathcal{L}_{\text{STPR}}$ show similar learning effects with $\mathcal{L}$, as individually presented in Fig. 6.13 and 6.14. This validates our intuition that the graph semantics learning during skeleton reconstruction (STPR) and the discriminative feature learning in the supervised contrastive learning (GPC) are compatible and they can be combined to facilitate the model training. To provide a further analysis of the learned skeleton representations, we follow [61] to estimate the *mean intra-class tightness (mACT)* and *mean inter-class looseness (mRCL)* of the learned skeleton graph representations *w.r.t.* the ground-truth classes. The mACT and mRCL can serve as effective evaluation metrics of the contrastive representation learning and identity-associated semantics learning[4]. As shown in Fig. 6.15 and 6.16, the training of our approach progressively and significantly improves

[4]According to the criterion in [61], a good model should satisfy: The same-class representations are gathered closer (higher mACT) while different-class representations possess larger distances (higher mRCL).

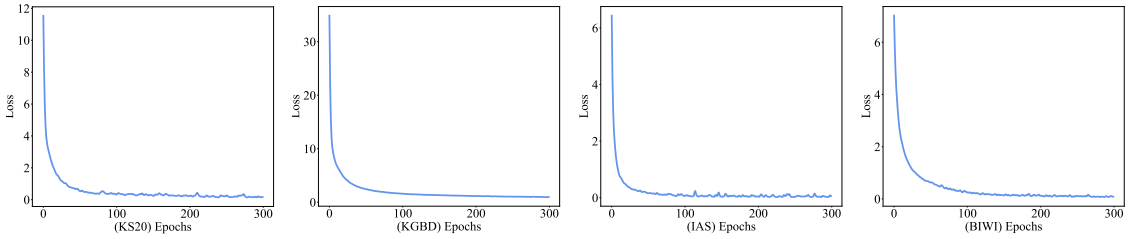


FIGURE 6.12: The total training loss curves on different training datasets.

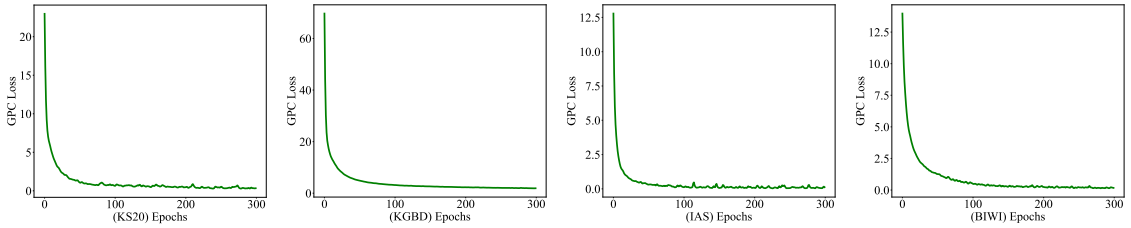


FIGURE 6.13: The graph prototype contrastive learning loss ($\mathcal{L}_{\text{GPC}}$) curves on different training datasets.

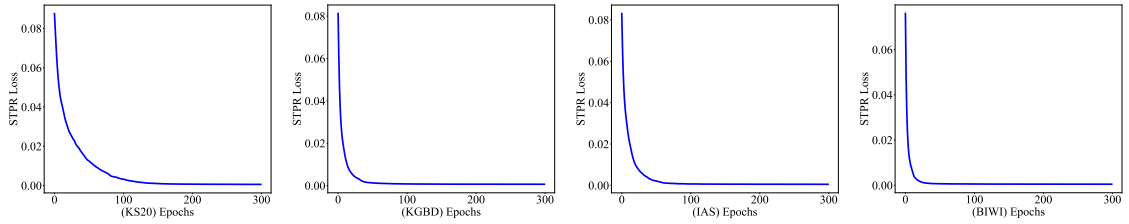


FIGURE 6.14: The graph structure-trajectory prompted reconstruction loss ($\mathcal{L}_{\text{STPR}}$) curves on different training datasets.

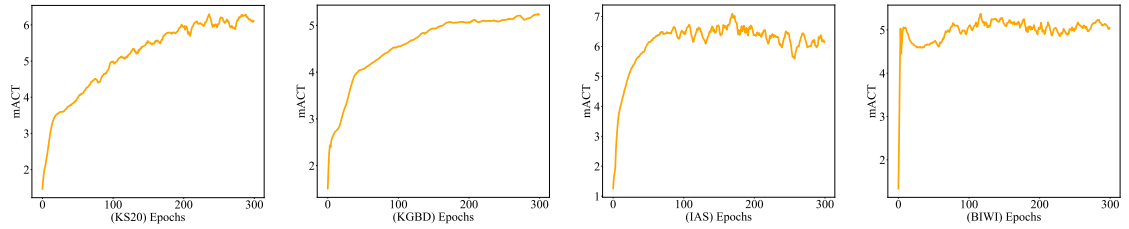


FIGURE 6.15: The mean intra-class tightness (mACT) of skeleton representations learned by our approach on different training datasets.

both mACT and mRCL of the learned skeleton graph representations on different datasets, which demonstrates that the proposed TranSG can encourage the model to capture effective class-related semantics (*e.g.*, inter-class differences) to learn more discriminative skeleton representations for person re-ID.

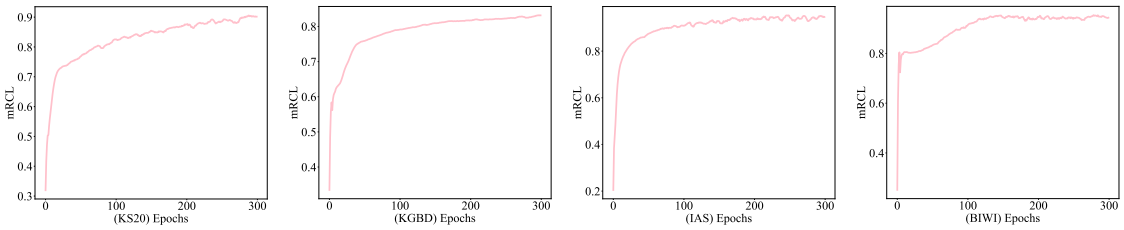


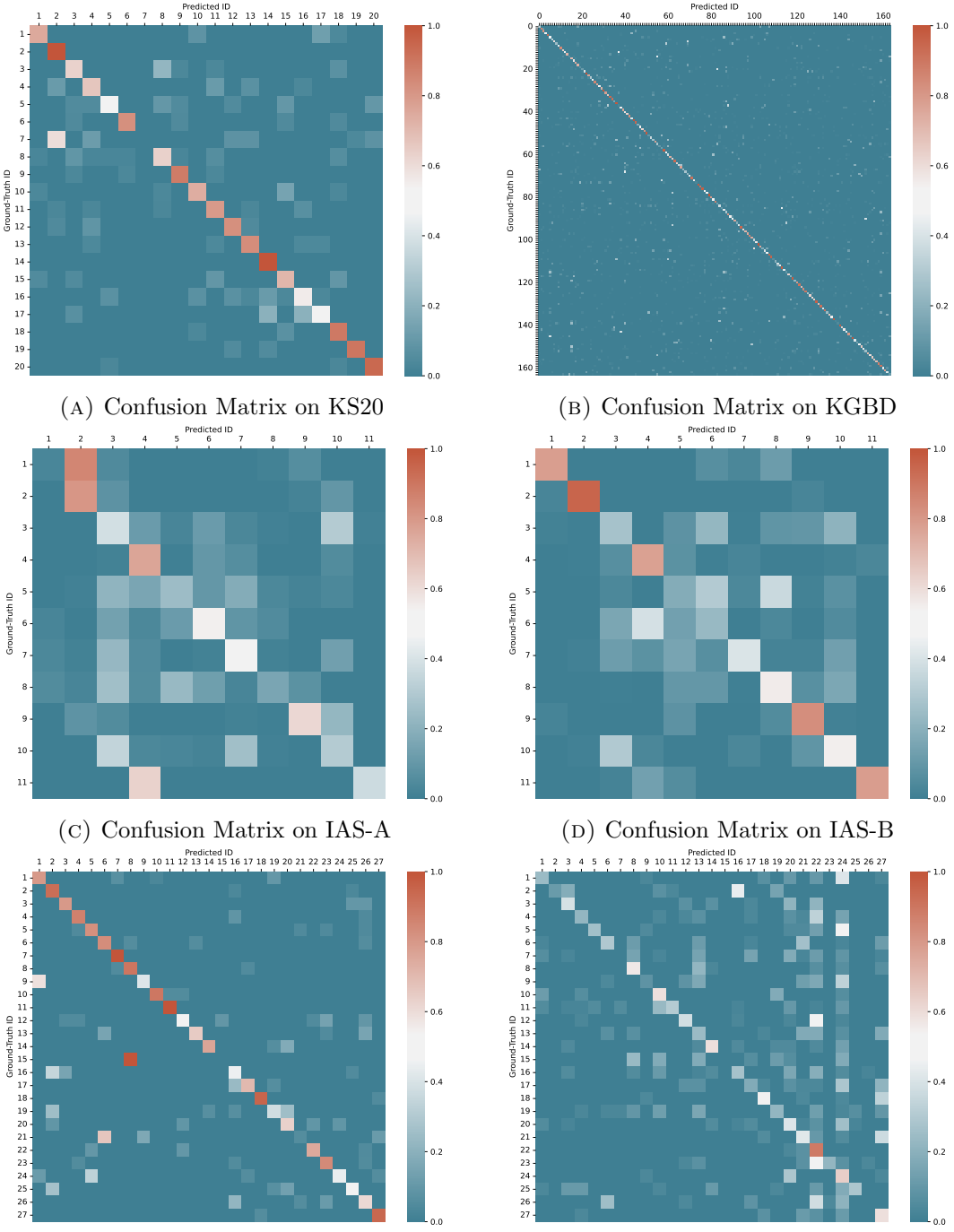
FIGURE 6.16: The mean inter-class looseness (mRCL) of skeleton representations learned by our approach on different training datasets.

6.4.12 Analysis of Confusion Matrix

As shown in Fig. 6.17, we visualize the confusion matrices of our approach when performing person re-ID with the Rank-1 matching (*i.e.*, predicting the identity of each probe sequence using the Rank-1 gallery sequence that has the smallest Euclidean distance) on all testing sets (probe sets). Fig. 6.17 (a)-(f) show that each confusion matrix possesses an evident alignment between the predicted identities and the ground-truth identities on the diagonal line. This suggests that skeleton sequences in most classes can be correctly matched between the probe set and gallery set in each dataset. Moreover, it can be seen that the ratios of classes with high accuracy (*i.e.*, ratios of red grids on the diagonal line) in KS20 and BIWI-Still are larger than that in IAS-A, IAS-B, KGBD, and BIWI-Walking. The larger numbers of white and red grids diffused *around* the diagonal lines, which represent the higher proportions of false matches, on the matrices of IAS-A (see Fig. 6.17 (c)) and BIWI-Walking (see Fig. 6.17 (f)) imply that our model tends to confuse skeleton sequences of more different identities on these datasets. These results are consistent with the performance results shown in our work.

6.5 Summary

In this chapter, we propose TranSG to learn effective representations from skeleton graphs for person re-ID. We devise a skeleton graph transformer (SGT) to perform full-relation learning of body-joint nodes to aggregate key body and motion features into graph representations. A graph prototype contrastive learning (GPC) approach is proposed to learn discriminative graph representations by contrasting their inherent similarity with the most representative graph features. Furthermore,



(A) Confusion Matrix on KS20

(B) Confusion Matrix on KGBD

(C) Confusion Matrix on IAS-A

(D) Confusion Matrix on IAS-B

(E) Confusion Matrix on BIWI-Still

(F) Confusion Matrix on BIWI-Walking

FIGURE 6.17: Visualization of confusion matrices on KS20 (a), KGBD (b), IAS-A (c), IAS-B (d), BIWI-Still (e), and BIWI-Walking (f) when using the Rank-1 matching. Note that abscissa and ordinate denote the predicted and ground-truth identities, respectively. The position in the a^{th} row and b^{th} column indicates that the testing samples belonging to the a^{th} identity is predicted as the b^{th} identity, while the corresponding value is the proportion of such samples to the same-identity samples in the testing set.

we design a graph structure-trajectory prompted reconstruction (STPR) mechanism to encourage learning richer graph semantics and key patterns for person re-ID. TranSG outperforms existing state-of-the-art models, and can be scalable to be applied to different scenarios.

Chapter 7

Skeleton-Based Person Re-ID Enhanced by Feature Re-Ranking

7.1 Introduction

3D skeletons (*i.e.*, 3D coordinates of body joints) captured by depth sensors such as Kinect [22] have been utilized as efficient body representations for person re-ID [2, 4, 24]. The skeleton-based person re-ID can be viewed as a *retrieval* problem to search skeleton sequence representations with the target identity in the gallery (*i.e.*, ranking list of representations) when given a probe skeleton sequence representation [5]. During the exploring and designing process of skeleton-based person re-ID models, researchers noticed that the quality of ranking of skeleton features could influence the final result of person re-ID, which suggests that a higher-quality ranking process can be integrated to improve the retrieval results and model performance (corresponding to the fourth challenge in Sec. 1.2). To this end, a good practice is to add a re-ranking process, which can synergize different metrics and context information to re-sort feature distances to improve retrieval accuracy [127, 150, 151]. However, recent research endeavors in skeleton-based person re-ID (see Chapter 4 and 5) [2, 4, 25] neither explicitly investigate existing feature re-ranking techniques nor explore their effectiveness on skeleton-based features or models, which motivates us to study a general re-ranking technique for this area.

This chapter has been published as: Haocong Rao, Yuan Li, and Chunyan Miao, "Revisiting k-Reciprocal Distance Re-ranking for Skeleton-Based Person Re-Identification," *IEEE Signal Processing Letters (SPL)*, 2022 [149]. DOI: 10.1109/LSP.2022.3212634.

Existing re-ranking models [127, 152, 153] are typically designed for conventional person re-ID methods that utilize visual features (*e.g.*, RGB-based appearances), while re-ranking for skeleton-based person re-ID still remains to be explored. Due to the modality gap between RGB images and skeletons, the RGB-based re-ranking paradigms using a certain distance metric (*e.g.*, Mahalanobis distance [127]) might be inapplicable to skeleton features, thus a more general feature re-ranking model is desired. On the other hand, different from traditional models [154–156] that leverage a single image for retrieval, skeleton-based person re-ID [5] generally matches identities between two skeleton sequences, each of which contains multiple skeleton representations that need to be integrated for better feature ranking.

To address the above challenges, we *for the first time* propose a general re-ranking method for skeleton-based person re-ID, which can exploit the salient features of skeleton sequences to perform k-reciprocal distance encoding for feature re-ranking. Specifically, considering that each skeleton representation is equally important and can individually characterize the identity [5], we first model skeleton representations within a sequence as a feature set [157], and propose the skeleton sequence pooling (SSP) that combines average pooling and max pooling to aggregate the most salient features of a skeleton sequence for better ranking. Then, we encode the neighbor (*i.e.*, k-reciprocal nearest neighbors [127]) distance and context information of skeleton sequence representations into a feature vector (defined as *"k-reciprocal distance vector"*) with the Jaccard metric, and exploit the fusion of both original Euclidean distance and k-reciprocal distance for re-ranking. Lastly, we devise the context-based Rank-1 voting, which jointly utilizes the local context (*i.e.*, top-k candidates) of gallery representations in both initial ranking list and re-ranking list to select the Rank-1 candidate for person re-ID.

With this chapter, we make the following contributions:

- We propose a generic re-ranking method with skeleton sequence pooling (SSP) that can aggregate salient features of a skeleton sequence for person re-ID re-ranking.
- We encode k-reciprocal nearest neighbors of skeleton sequence representations into k-reciprocal distances, which is fused with the original distance to re-rank features.

- We devise the context-based Rank-1 voting that jointly exploits the initial ranking and re-ranking lists to vote for a more reliable Rank-1 candidate for person re-ID.
- Experiments on three benchmarks show that our method is highly effective on re-ranking various state-of-the-art skeleton representations to improve their performance.

7.2 The Proposed Approach

In this section, we formulate the definition of skeleton-based person re-ID re-ranking, and present the details of each component in our method, whose overview is given in Fig. 7.1.

7.2.1 Problem Definition

Suppose that a 3D skeleton sequence is $\boldsymbol{S}_{1:f} = (\boldsymbol{S}_1, \cdots, \boldsymbol{S}_f) \in \mathbb{R}^{f \times X}$, where $\boldsymbol{S}_i \in \mathbb{R}^X$ denotes the i^{th} skeleton with 3D coordinates of J body joints and $X = 3 \times J$. A skeleton-based person re-ID model encodes f skeletons $\boldsymbol{S}_{1:f}$ into effective representations $\boldsymbol{v}_{1:f} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_f)$, which are then integrated into a sequence representation by $P(\boldsymbol{v}_{1:f}) = \overline{\boldsymbol{v}}$, where $P(\cdot)$ is a transformation function and $\overline{\boldsymbol{v}}$ denotes the final sequence representation for person re-ID.

The goal of skeleton-based person re-ID is to search skeleton sequence representations of the same person in the gallery set when given a representation from the probe set. Formally, given a skeleton sequence representation $\boldsymbol{p}$ of the person P from the probe set and n representations from the gallery set, denoted as $\mathcal{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_n\}$, the distance between two representations is measured by Euclidean distance [5] with:

$$D_E(\boldsymbol{p}, \boldsymbol{g}_i) = \|\boldsymbol{p} - \boldsymbol{g}_i\|_2 , \tag{7.1}$$

where $D_E(\boldsymbol{p}, \boldsymbol{g}_i)$ denotes the Euclidean distance between the probe representation $\boldsymbol{p}$ and gallery representation $\boldsymbol{g}_i$ and $\|\cdot\|_2$ represents the l_2 norm. Based on the Euclidean distance in Eq. (7.1), we can obtain the *initial* ranking list $L^0(\boldsymbol{p}, \mathcal{G}) =$

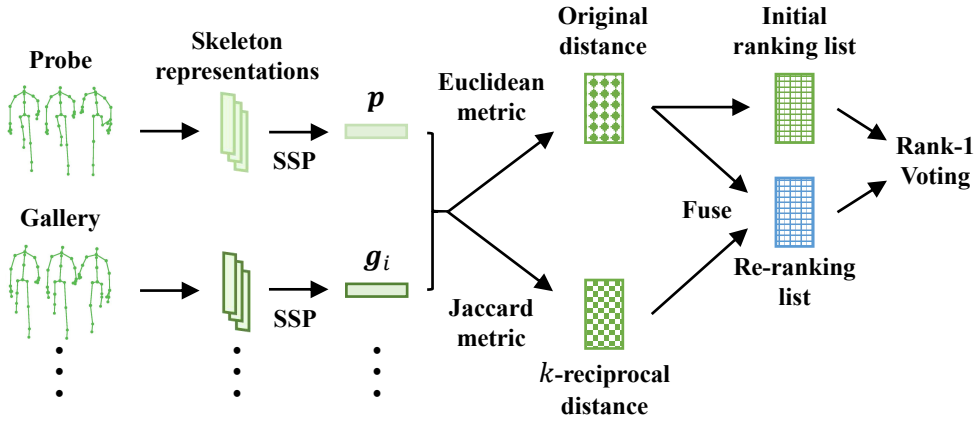


FIGURE 7.1: Overview of our method with skeleton sequence pooling (SSP), fused distance based re-ranking, and Rank-1 voting for skeleton-based person re-ID.

$\{\boldsymbol{g}_1^0, \boldsymbol{g}_2^0, \cdots, \boldsymbol{g}_n^0\}$, where $D_E(\boldsymbol{p}, \boldsymbol{g}_i^0) \leq D_E(\boldsymbol{p}, \boldsymbol{g}_{i+1}^0)$. The goal of our approach is to re-rank $L(\boldsymbol{p}, \mathcal{G})$, such that more representations with the target identity are listed as candidates on the top to improve the person re-ID performance.

7.2.2 Skeleton Sequence Pooling

In the skeleton-based re-ID task, a model learns representations for *multiple* skeletons in a sequence (see Sec. 7.2.1) to characterize *structured* body features (*i.e.*, body joints) and gait patterns, which fundamentally differs from conventional person re-ID methods that utilize a *single* RGB image to match persons. To aggregate the most salient features from the learned skeleton representations $\boldsymbol{v}_{1:f} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_f)$ within a sequence for ranking, we propose to regard $\boldsymbol{v}_{1:f}$ as a feature set [157] (*i.e.*, assuming that each skeleton representation can effectively characterize the identity), and apply the *average pooling (AP)* and *max pooling (MP)* on the set to obtain the corresponding sequence representation by:

$$\begin{aligned} \overline{\boldsymbol{v}} &= P(\boldsymbol{v}_{1:f}) = \text{average}(\boldsymbol{v}_{1:f}) + \max(\boldsymbol{v}_{1:f}) \\ &= \left(\frac{1}{f}\sum_{t=1}^{f}[\boldsymbol{v}_t^1, \cdots, \boldsymbol{v}_t^n]\right) + [\boldsymbol{v}_{x_1}^1, \cdots, \boldsymbol{v}_{x_n}^n], \end{aligned} \tag{7.2}$$

where $\text{average}(\cdot)$, $\max(\cdot)$ are element-based AP and MP operations over the set dimension, $\overline{\boldsymbol{v}} \in \mathbb{R}^n$ denotes the final sequence representation, $\boldsymbol{v}_i = [\boldsymbol{v}_i^1, \cdots, \boldsymbol{v}_i^n] \in \mathbb{R}^n$, n is the dimension number of skeleton representations, $\boldsymbol{v}_i^n \in \mathbb{R}$ is the n^{th}

element of the i^{th} skeleton representation (*i.e.*, feature vector $\boldsymbol{v}_i$), and we have $x_n = \underset{i}{\text{argmax}}\, \boldsymbol{v}_i^n$. AP and MP can be viewed as unique statistical functions [49, 157] to extract the global average information and salient features of all skeletons within a sequence. As both addition and concatenation achieve similar performance in practice (see Sec. 7.4.1), we adopt addition in Eq. (7.2) to retain a small feature dimension (*i.e.*, half of dimension compared with concatenation) to reduce the computational cost. Note that we employ skeleton sequence pooling in both probe and gallery sets, where $\overline{\boldsymbol{v}}$ is denoted as $\boldsymbol{p}$ and $\boldsymbol{g}$, respectively, for simplicity of presentation.

7.2.3 k-Reciprocal Distance Encoding

For a probe representation $\boldsymbol{p}$, we define its *k-nearest neighbors (k-NN)* as the top-k gallery representations of the initial ranking list [158], which can be represented as:

$$\mathcal{N}(\boldsymbol{p}, k) = \left\{\boldsymbol{g}_1^0, \boldsymbol{g}_2^0, \cdots, \boldsymbol{g}_k^0\right\}, \tag{7.3}$$

where $|\mathcal{N}(\boldsymbol{p}, k)| = k$ and $|\cdot|$ represents the number of candidates in the set. Following [127], we define the k-reciprocal nearest neighbors $\mathcal{R}(\boldsymbol{p}, k)$ by:

$$\mathcal{R}(\boldsymbol{p}, k) = \left\{\boldsymbol{g}_i \mid (\boldsymbol{g}_i \in \mathcal{N}(\boldsymbol{p}, k)) \wedge (\boldsymbol{p} \in \mathcal{N}(\boldsymbol{g}_i, k))\right\}. \tag{7.4}$$

The k-reciprocal nearest neighbors $\mathcal{R}(\boldsymbol{p}, k)$ contain more related representations of $\boldsymbol{p}$ than the k-NN set $\mathcal{N}(\boldsymbol{p}, k)$, as it takes the intersection of both original set and the k-NN sets of its neighbors. Different from [127] that performs an expansion process to add more candidates from $\mathcal{R}(\boldsymbol{g}_i, \frac{1}{2}k)$ to the set, where $\boldsymbol{g}_i \in \mathcal{R}(\boldsymbol{p}, k)$, we adopt the original set $\mathcal{R}(\boldsymbol{p}, k)$ that is stricter against including too many negative candidates.

Then, we encode the k-reciprocal nearest neighbor set into a *k-reciprocal feature vector* $\mathcal{F}_{\boldsymbol{p}} = \left[\mathcal{F}_{\boldsymbol{p},\boldsymbol{g}_1}, \mathcal{F}_{\boldsymbol{p},\boldsymbol{g}_2}, \cdots, \mathcal{F}_{\boldsymbol{p},\boldsymbol{g}_n}\right]$, where each item of the vector is computed by the Gaussian kernel as:

$$\mathcal{F}_{\boldsymbol{p},\boldsymbol{g}_i} = \begin{cases} \exp\left(-D_E\left(\boldsymbol{p}, \boldsymbol{g}_i\right)\right) & \text{if } g_i \in \mathcal{R}(\boldsymbol{p}, k) \\ 0 & \text{otherwise.} \end{cases} \tag{7.5}$$

$\mathcal{F}_{\boldsymbol{p},\boldsymbol{g}_i}$ can be viewed as a distance-based indicator to show the inherent similarity between the probe representation $\boldsymbol{p}$ and each of its k-reciprocal nearest neighbor, *i.e.*, the neighbor with smaller original distance is more similar and assigned with a larger weight. In this way, the vector $\mathcal{F}_{\boldsymbol{p}}$ combines both contextual knowledge of neighbors (see Eq. (7.4)) and distance information (see Eq. (7.1)) to characterize the discriminative relationship between $\boldsymbol{p}$ and $\mathcal{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_n\}$.

Considering that the overlapping degree of k-reciprocal nearest neighbor sets between two skeleton representations can indicate their inherent correlations [127, 159], *i.e.*, representations that share more neighbors in the sets possess higher similarity, we exploit the Jaccard metric to re-compute the new distance, *i.e.*, *k-reciprocal distance*, between $\boldsymbol{p}$ and $\boldsymbol{g}_i$:

$$\begin{aligned} D_J\left(\boldsymbol{p}, \boldsymbol{g}_i\right) &= 1 - \frac{\left|\mathcal{R}(\boldsymbol{p}, k) \cap \mathcal{R}\left(\boldsymbol{g}_i, k\right)\right|}{\left|\mathcal{R}(\boldsymbol{p}, k) \cup \mathcal{R}\left(\boldsymbol{g}_i, k\right)\right|} \\ &= 1 - \frac{\sum_{j=1}^{N} \min\left(\mathcal{F}_{\boldsymbol{p},\boldsymbol{g}_j}, \mathcal{F}_{\boldsymbol{g}_i,\boldsymbol{g}_j}\right)}{\sum_{j=1}^{N} \max\left(\mathcal{F}_{\boldsymbol{p},\boldsymbol{g}_j}, \mathcal{F}_{\boldsymbol{g}_i,\boldsymbol{g}_j}\right)}, \end{aligned} \tag{7.6}$$

where the numbers of candidates in the interaction and union of sets are calculated by $\left\|\min\left(\mathcal{F}_{\boldsymbol{p}}, \mathcal{F}_{\boldsymbol{g}_i}\right)\right\|_1 = \sum_{j=1}^{N} \min\left(\mathcal{F}_{\boldsymbol{p},\boldsymbol{g}_j}, \mathcal{F}_{\boldsymbol{g}_i,\boldsymbol{g}_j}\right)$ and $\left\|\max\left(\mathcal{F}_{\boldsymbol{p}}, \mathcal{F}_{\boldsymbol{g}_i}\right)\right\|_1 = \sum_{j=1}^{N} \max\left(\mathcal{F}_{\boldsymbol{p},\boldsymbol{g}_j}, \mathcal{F}_{\boldsymbol{g}_i,\boldsymbol{g}_j}\right)$, respectively, based on the definition of Eq. (7.5). $\min(\cdot, \cdot)$ and $\max(\cdot, \cdot)$ denote the operation of the element-based minimization and maximization. $\|\cdot\|_1$ is the l_1 norm. Eq. (7.6) transforms the set comparison under the Jaccard metric into pure vector calculation, which enables easily computing the k-reciprocal distance between two representations.

Local Query Expansion. As the k-nearest neighbors of the probe representation $\boldsymbol{p}$ share highly similar features, we can exploit them to implement a local expansion of query [127]. Specifically, we average the k-reciprocal distance features of the neighbors of $\boldsymbol{p}$ to define the corresponding query as:

$$\mathcal{F}_{\boldsymbol{p}} = \frac{1}{|\mathcal{N}(\boldsymbol{p}, k)|} \sum_{\boldsymbol{g}_i \in \mathcal{N}(\boldsymbol{p},k)} \mathcal{F}_{\boldsymbol{g}_i}. \tag{7.7}$$

This local query expansion is performed on both probe representation $\boldsymbol{p}$ and gallery representation $\boldsymbol{g}_i$. For clarity, we denote the sizes of $\mathcal{R}(\boldsymbol{p}, k)$ in Eq. (7.5) and

$\mathcal{N}(\boldsymbol{p}, k)$ in Eq. (7.7) as k_1 and k_2, respectively. To reduce the potential noisy in the k-nearest neighbors, we follow [127] to set a small value of k_2.

7.2.4 Fused Distance Based Re-Ranking

We re-rank the initial ranking list $L^0(\boldsymbol{p}, \mathcal{G})$ based on the fused distance D^*, combining both original Euclidean distance D_E (Eq. (7.1)) and k-reciprocal distance D_J (Eq. (7.6)) with:

$$D^*\left(\boldsymbol{p}, \boldsymbol{g}_i\right) = \beta D_E\left(\boldsymbol{p}, \boldsymbol{g}_i\right) + (1-\beta) D_J\left(\boldsymbol{p}, \boldsymbol{g}_i\right), \tag{7.8}$$

where $\beta \in [0, 1]$ is the fusion coefficient. The larger value of β indicates higher influence of the original distance on the re-ranking, while $\beta = 0$ means only considering the k-reciprocal distance. We denote the re-ranking list based on the fused distance in Eq. (7.8) as $L^*(\boldsymbol{p}, \mathcal{G}) = \{\boldsymbol{g}_1^*, \boldsymbol{g}_2^*, \cdots, \boldsymbol{g}_n^*\}$.

7.2.5 Context-Based Rank-1 Voting

To further improve the accuracy of Rank-1 matching, we propose to simultaneously maintain both original ranking list (L^0) and re-ranking list (L^*) to vote for the nearest neighbor (*i.e.*, Rank-1 candidate) of the probe representation $\boldsymbol{p}$ based on the context of lists. In particular, when the Rank-1 candidate is different between two lists, we give the preference to the one that has more same-identity candidates in the *local* context (*i.e.*, first k candidates) of the ranking list, formulated as:

$$\boldsymbol{g}_1 = \begin{cases} \boldsymbol{g}_1^* & \text{if } |\boldsymbol{g}_i^* \in L^*(\boldsymbol{p}, \mathcal{G})| \geq |\boldsymbol{g}_i^0 \in L^0(\boldsymbol{p}, \mathcal{G})|, \\ & \text{where } \mathrm{id}(\boldsymbol{g}_i^*) = \mathrm{id}(\boldsymbol{g}_1^*),\ \mathrm{id}(\boldsymbol{g}_i^0) = \mathrm{id}(\boldsymbol{g}_1^0). \\ \boldsymbol{g}_1^0 & \text{otherwise.} \end{cases} \tag{7.9}$$

In Eq. (7.9), $\boldsymbol{g}_1$ denotes the final Rank-1 gallery candidate with higher votes of the same identity ($\mathrm{id}(\cdot)$) in the first k_3 positions of lists. For convenience, we use $L^*(\boldsymbol{p}, \mathcal{G})$ and $L^0(\boldsymbol{p}, \mathcal{G})$ to represent $\{\boldsymbol{g}_1^*, \boldsymbol{g}_2^*, \cdots, \boldsymbol{g}_{k_3}^*\}$ and $\{\boldsymbol{g}_1^0, \boldsymbol{g}_2^0, \cdots, \boldsymbol{g}_{k_3}^0\}$, respectively. As the context of top-ranking representations typically contain multiple related representations with the same identity in the ranking list, the proposed

context-based Rank-1 voting can jointly exploit two ranking lists with the available gallery identity information to select a more reliable Rank-1 candidate to perform person re-ID. It is worth noting that the Rank-1 voting is only utilized for Rank-1 matching while NOT changing the re-ranking list and corresponding mAP results.

7.3 Experiments

7.3.1 Experimental Setup

7.3.1.1 Datasets

Our approach is evaluated on three public benchmarks: *IAS-Lab RGBD-ID Dataset (IAS)* [120], *KS20 VisLab Multi-View Kinect Skeleton Dataset (KS20)* [119], and *BIWI RGBD-ID Dataset (BIWI)* [1], which contain skeletons of 11, 20, and 50 different persons, respectively. We follow the commonly-used probe and gallery settings in [5]. The full description and visual samples of datasets are provided in Sec. 4.3.1.1.

7.3.1.2 Probe and Gallery Settings

We follow the commonly-used settings of probe and gallery in the literature [5]: For the BIWI and IAS datasets, as different testing sets are non-overlapped and contain all pedestrians under different scenes, we evaluate our approach on each testing set by setting it as the probe while the other one is adopted as the gallery. The KS20 dataset collects skeleton data of pedestrians from five different viewpoints, including 0°, 30°, 90°, 130°, and 180°. We randomly take one skeleton sequence from each view as the probe sequence and use one half of the remaining sequences for training and the other half as the gallery. Experiments with each setup are repeated for multiple times and the average performance is reported in this work.

7.3.1.3 Implementation Details

Dataset Preprocessing Setups. To avoid ineffective skeleton recording, we discard the first and last 10 skeleton frames of each original skeleton sequence. For KS20, BIWI, and IAS datasets, all skeleton sequences are normalized by subtracting the spine joint position from each joint of the same skeleton so that the skeleton is translation invariant [124]. Then, we spilt all normalized skeleton sequences in the training sets into multiple shorter skeleton sequences (*i.e.*, $\boldsymbol{S}_{1:f}$) with length f by a step of $\frac{f}{2}$, which aims to obtain as many 3D skeleton sequences as possible to train our approach. We split all skeleton sequences in the gallery and probe sets into shorter and non-overlapping sequences with length f. Unless explicitly specified, the skeleton sequence $\boldsymbol{S}_{1:f}$ in our work refers to those split and normalized sequences used in learning, rather than those original skeleton sequences provided by datasets. We follow the data augmentation strategy used in [4, 7, 8] to sample more sequences for different identities in the training set, and train our approach with randomly shuffled and unlabeled skeleton sequences.

Model Parameter Setups. The body-joint numbers are $J = 25$ for KS20, $J = 20$ for BIWI and IAS. We follow [5] to set the sequence length $f = 6$ for a fair comparison and select the best model parameters for existing skeleton-based person re-ID methods (SGELA [4], MG-SCR [7], SM-SGE [8], SimMC [5]). We follow [127] to set $k_1 = 20$, $k_2 = 6$, and $\beta = 0.3$ for the original k-reciprocal re-ranking (k-RR) method. For our approach, we empirically employ $k_1 = 30$, $k_2 = 6$, and $k_3 = 4$ on all datasets. The β is set to 0.7 for SGELA, 0.8 for SimMC, and 0.6 for MG-SCR and SM-SGE.

Baseline Method Setups. To validate the effectiveness and generality of the proposed re-ranking approach, we apply it to four state-of-the-art skeleton-based person re-ID models [4, 5, 7, 8], and compare the person re-ID performance after re-ranking with the original baseline performance. Our approach is also compared with the original k-reciprocal re-ranking method k-RR [127] on different datasets.

7.3.1.4 Evaluation Metrics

Rank-1 accuracy (R_1) and Mean Average Precision (mAP) [49] are adopted to quantitatively evaluate the performance of person re-ID.

TABLE 7.1: Original and re-ranking performance of skeleton-based person re-ID models on different datasets. Our method is compared with the original k-reciprocal distance re-ranking (k-RR) [127]. The bold indicates better re-ranking performance.

Methods	KS20		IAS-A		IAS-B		BIWI-S		BIWI-W	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
SGELA [4]	45.0	21.2	16.7	13.1	22.2	14.0	25.8	15.1	11.7	19.0
SGELA + k-RR [127]	35.0	20.0	21.9	13.2	16.0	14.0	36.7	17.7	10.5	19.8
SGELA + Ours	**47.0**	**22.0**	**22.0**	**13.5**	**22.4**	**14.5**	**40.5**	**18.6**	**11.9**	**20.1**
MG-SCR [7]	46.3	10.4	36.4	14.1	32.4	12.9	20.1	7.6	10.8	11.9
MG-SCR + k-RR	36.3	9.4	30.4	**14.3**	30.2	**13.9**	21.7	**8.0**	9.1	11.6
MG-SCR + Ours	**47.9**	**10.4**	**37.3**	14.1	**36.3**	**13.9**	**22.3**	7.1	9.4	11.1
SM-SGE [8]	45.9	9.5	34.0	13.6	38.9	13.3	31.3	10.1	13.2	15.2
SM-SGE+ k-RR	33.6	8.7	31.9	**14.4**	31.1	13.8	**31.4**	**10.8**	12.7	**15.4**
SM-SGE + Ours	**46.9**	8.9	**36.9**	14.3	**42.6**	**13.9**	30.3	10.3	**13.7**	14.4
SimMC [5]	66.4	22.3	44.8	18.7	46.3	22.9	41.7	12.3	24.5	19.9
SimMC + k-RR	63.3	23.2	44.1	**23.4**	41.9	26.5	35.7	**15.5**	19.1	**22.6**
SimMC + Ours	**69.5**	**24.0**	**55.1**	**23.4**	**49.3**	**27.2**	**48.6**	14.9	**25.4**	20.5

7.3.2 Experimental Results

As shown in Table 7.1, the proposed re-ranking approach consistently improves the performance of the original SGELA [4], MG-SCR [7], and SimMC [5] models by up to 5.3% for Rank-1 accuracy and 4.7% for mAP on KS20, IAS-A, and IAS-B testing sets. This demonstrates the effectiveness of our approach when applied to different models. Although the k-RR model [127] obtains slightly better mAP results on BIWI-S and BIWI-W, it performs poorly on many other datasets and even largely degrades the performance of original models, which implies that the k-RR might lack the generality to be applied to re-ranking skeleton features. In contrast, our approach is more stable with higher performance than k-RR by 5.8% for Rank-1 accuracy and 0.05% for mAP in average on all datasets, which suggests that our method that combines skeleton sequence pooling (SSP) and fused distance based re-ranking is more effective on skeleton-based models.

7.4 Further Analysis

7.4.1 Ablation Study

As reported in Table 7.2, using the fused distance based re-ranking (FDR) achieves higher person re-ID performance than the baseline by 1.2-4.1% mAP, while solely

TABLE 7.2: Ablation study of our approach (on SimMC) with different configurations: Direct fused distance based re-ranking (FDR), context-based Rank-1 voting (RV), and skeleton sequence pooling (SSP). We compare the addition (Add.) and concatenation (Concat.) of two pooling features for SSP.

	KS20		IAS-A	
Configurations	**R_1**	**mAP**	**R_1**	**mAP**
Baseline (SimMC)	66.4	22.3	44.8	18.7
+ FDR	66.4	23.5	44.9	22.8
+ FDR + RV	68.8	23.5	53.9	22.8
+ SSP (Add.) + FDR	68.2	24.0	45.0	23.4
+ SSP (Concat.) + FDR + RV	69.5	23.8	54.7	22.9
+ SSP (Add.) + FDR + RV	69.5	24.0	55.1	23.4

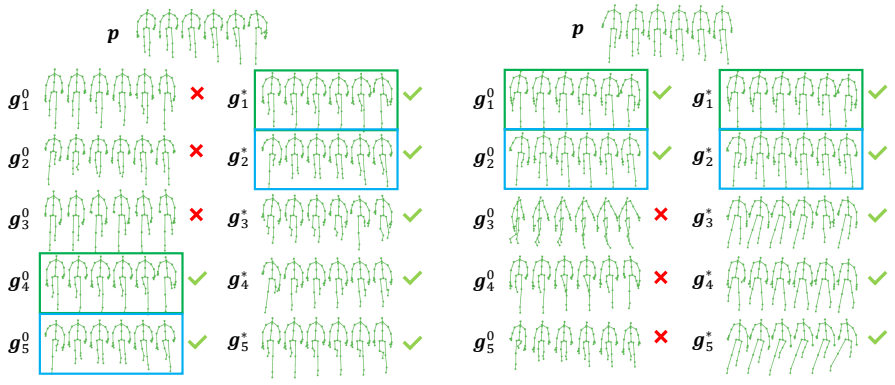


FIGURE 7.2: Result comparison between the original ranking list ($\boldsymbol{g}_i^0$) and re-ranking list ($\boldsymbol{g}_i^*$) when applying our approach to SimMC on two different probe sequences ($\boldsymbol{p}$) of KS20. ✓ and ✗ indicate the true match and false match. The proposed re-ranking can assign higher ranks to correct gallery sequences (green and blue boxes) and add more true matches to the top list.

adding SSP or the proposed Rank-1 voting (RV) can improve Rank-1 accuracy by 2.4-9.0% and 0.1-1.8%, respectively, which demonstrates their effectiveness in our approach. Combining both of them achieves the best Rank-1 accuracy and mAP on different datasets, as they enables our model to aggregate salient skeleton sequence features (SSP) for ranking while improving the performance of Rank-1 matching with Rank-1 voting (RV). The results also show that different combining manners (Add. or Concat.) in SSP (see Sec. 7.2.2) achieve similar performance. Other datasets also report similar results.

7.4.2 Qualitative Analysis

To validate the effectiveness of our approach, we visualize skeleton sequences in the probe and their corresponding original ranking and re-ranking lists of gallery

TABLE 7.3: Performance of our approach when applied to the SimMC model with different values of k_1 (used in $\mathcal{R}(\boldsymbol{p}, k_1)$) on KS20. We adjust this parameter while keeping others consistent.

k_1	**5**	**10**	**15**	**20**	**25**	**30**
$\mathbf{R}_1$	69.6	69.1	68.6	69.5	69.1	69.5
mAP	24.1	24.4	24.3	24.3	24.1	24.0

TABLE 7.4: Performance of our approach when applied to the SimMC model with different values of k_2 (used in query expansion) on KS20. We adjust this parameter while keeping others consistent.

k_2	**2**	**4**	**6**	**8**	**10**	**12**
$\mathbf{R}_1$	69.7	69.9	69.5	68.9	69.5	69.1
mAP	23.9	24.1	24.0	23.7	23.5	23.3

TABLE 7.5: Performance of our approach when applied to the SimMC model with different values of k_3 (used in Rank-1 voting) on KS20. We adjust this parameter while keeping others consistent.

k_3	**4**	**8**	**12**	**16**	**20**	**24**
$\mathbf{R}_1$	69.5	67.8	68.0	67.6	68.2	68.0
mAP	24.0	24.0	24.0	24.0	24.0	24.0

TABLE 7.6: Performance of our approach when applied to the SimMC model with different values of β (distance fusion coefficient) on KS20. We adjust this parameter while keeping others consistent.

β	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
$\mathbf{R}_1$	70.1	70.3	70.1	69.7	69.5	65.8
mAP	21.4	23.4	23.6	23.8	24.0	22.8

sequences on KS20 dataset. As presented in Fig. 7.2, compared with the original list that assigns high ranks to false matches, our re-ranking approach can move the nearest true matches of gallery sequences to the top and replace the false ones. Besides, when two lists share true matches on the top, the re-ranking list can keep those consistent sequences while adding more related sequences of the same identity to the list. These results show the validity of our approach on re-ranking skeleton sequence representations for better person re-ID.

7.4.3 Analysis of Hyperparameters

We evaluate the effects of different hyper-parameters on our approach in Table 6.7 and obtain the following observations and conclusions: **(1)** Our approach is not sensitive to different settings of k_1 and k_2 despite with slight performance variations.

In practice, we select the best value for our approach to achieve better average performance on all datasets. **(2)** Adopting a lower value of k_3 can significantly improve Rank-1 accuracy, which suggests that exploiting the topper candidates (*e.g.*, top four gallery representations ($k_3 = 4$)) as a context for voting can achieve more Rank-1 true matches. **(3)** Solely using the k-reciprocal distance for re-ranking ($\beta = 0.0$) obtains higher performance than using the original Euclidean distance ($\beta = 1.0$), while an appropriate combination of them can achieve the best overall performance. These results further demonstrate the necessity of employing the fused distance to re-rank skeleton representations for better person re-ID.

7.5 Summary

In this chapter, we revisit the k-reciprocal distance re-ranking and devise a generic re-ranking method for skeleton-based person re-ID. The skeleton sequence pooling is proposed to aggregate the salient skeleton features of a sequence for feature ranking. We fuse both original Euclidean distance and k-reciprocal distance of skeleton sequences to perform re-ranking. We further devise the context-based Rank-1 voting to combine both initial ranking and re-ranking lists to enhance the Rank-1 matching. Our approach can effectively re-rank existing skeleton features to improve person re-ID performance.

Chapter 8

Conclusions and Future Works

8.1 Conclusions

In this thesis, we focus on the problem of skeleton-based person re-ID, and explore AI-empowered solutions to effectively learn discriminative body structure features and motion patterns from skeleton data to identify different persons. With the aim to achieve more effective and efficient skeleton-based person re-ID, we concentrate on four key challenges in this area, including *unlabeled skeleton learning*, *multi-level body modeling*, *body relation learning*, *general model enhancement*, and respectively propose four novel AI-based models and frameworks to address them.

- To address the challenge of simultaneously capturing identity-related and general skeleton features from *unlabeled skeletons* (see Chapter 4), we devise an unsupervised Simple Masked Contrastive learning (SimMC) framework. It leverages unsupervised Masked Prototype Contrastive (MPC) learning to capture the most representative and discriminative skeleton features (prototypes) from masked skeleton sub-sequences, and exploits the Masked Intra-sequence Contrastive (MIC) learning to encourage capturing domain-general skeleton semantics such as motion continuity to enhance the skeleton representation learning. To the best of our knowledge, SimMC is the first unsupervised skeleton representation learning framework for skeleton-based person re-ID. Empirical evaluations on four benchmark datasets show that

SimMC significantly outperforms most state-of-the-art skeleton-based methods, and can serve as a generic contrastive paradigm to fine-tune and boost existing skeleton representations.

- To solve the problem of *multi-level body modeling* using skeletons (see Chapter 5), we devise an unsupervised Hierarchical skeleton Meta-Prototype Contrastive learning (Hi-MPC) approach, which exploits hierarchical skeleton representations from coarse to fine to characterize anthropometric and kinetic body features. To improve the robustness and consistency of prototype estimation and learning in previous studies (see Chapter 4), we introduce meta-prototype contrastive learning by constructing different homogeneous contrastive subspaces to jointly learn discriminative prototypes. Moreover, we propose the *first* Hard Skeleton Mining (HSM) mechanism specifically for skeleton-based person re-ID, which can focus on more important skeleton representations at different levels to enhance the model performance. Extensive experiments on five public benchmarks, including multi-view and RGB-estimated skeletons, demonstrate the effectiveness and scalability of Hi-MPC. We also reveal the feasibility of utilizing more concise and abstract skeleton representations to perform person re-ID, opening new possibilities for efficient skeletal modeling and model compression in future.

- To resolve the challenge of *body relation learning* within skeletons (see Chapter 6), we propose a generic Transformer-based Skeleton Graph prototype contrastive learning (TranSG) paradigm, which concurrently captures structural and actional relations of from both adjacent and non-adjacent body-component nodes. We further improve the reliability of previous skeleton prototype learning (see Chapter 4 and 5) by introducing the supervision of ground-truth labels. To capture general valuable body structural and motion features under the supervised prototype learning, we propose the Structure-Trajectory Prompted Reconstruction (STPR) mechanism to enhance the general high-level skeleton semantics learning for person re-ID. To the best of our knowledge, TranSG is the first transformer paradigm that unifies skeletal relation learning specifically for skeleton-based person re-ID. Comprehensive experiments on five public benchmarks demonstrate the effectiveness of TranSG and its high scalability to be applied to different-level graph modeling, RGB-estimated or unlabeled skeleton data.

- To tackle the challenge of *general model enhancement* (see Chapter 7), we leverage the k-reciprocal distance to design a general feature re-ranking framework specifically for skeleton-based person re-ID. To aggregate the most salient skeleton features and the most crucial neighbor context information for re-ranking, we propose the Skeleton Sequence Pooling (SSP) and context-based Rank-1 voting techniques, which are further combined with the fused Euclidean and k-reciprocal distance based feature re-ranking to significantly improve performance of existing skeleton-based person re-ID models. With the visualization of re-ranking results, we can intuitively view the top similar skeleton sequences and qualitatively validated the effectiveness of our approach. Experiments on different benchmarks show that our method is highly effective on re-ranking various state-of-the-art skeleton representations to improve their performance.

This thesis proposes the above AI-based solutions and provides in-depth empirical and theoretical analyses with beneficial insights. It involves the systematic design of skeleton-based person re-ID models tailored to diverse application scenarios, including multi-view skeletons, RGB-estimated skeletons, skeletons without labels, and skeletons with different topological structures. Based on thorough experiments, the effectiveness and generality of these models for various skeleton-based person re-ID tasks are validated, and we further demonstrate their superiority over previous methods in terms of model performance and efficiency.

Furthermore, our studies possess the potential to be applied to various real-world applications, with significant and positive impacts generated across multiple domains such as healthcare, security, and mobile applications.

8.2 Future Research Directions

There are several research directions for our future works. First, we systematically present three promising directions for model improvement, namely body-component relation learning (Sec. 8.2.1), skeleton sample augmentation (Sec. 8.2.2), and importance-aware intra-sequence learning (Sec. 8.2.3), which help devise a more effective AI-empowered approach for skeleton-based person re-ID. We also discuss interdisciplinary research directions such as skeleton-based models for

healthcare (Sec. 8.2.4) and other possible directions in our future works (Sec. 8.2.5).

8.2.1 Body-Component Relation Learning

An important future direction is to devise more effective body modeling mechanisms or skeleton semantics learning tasks to learn body-component relations and patterns. Firstly, although previous works have investigated different level body-component relations (*e.g.*, part-level and limb-level relations [7, 133]), their pre-defined scale and number of body components are usually fixed based on domain knowledge. Hence, to enhance the flexibility of body modeling, we could explore an adaptive body-component partitioning mechanism to automatically focus on the relations between fine-grained key components and coarse-grained body parts. Secondly, As there usually exist strong correlations between nearby body joints/components within a local spatial-temporal *context* [4], we can design self-supervised context-based semantic tasks to encourage learning inherent relations within body and motion. In our TranSG approach (see Chapter 6), we devise the structure-trajectory prompted reconstruction mechanism to learn both structural relations and motion relations of body joints via skeleton reconstruction. However, it relies on skeleton reconstruction without exploring other forms of tasks. It is therefore feasible to enhance the capture of temporal motion correlations by letting the model predict the future skeletons. Such tasks can also be applied to higher level body components such as representations of limbs to capture more comprehensive (*e.g.*, global) patterns.

On the other hand, most methods learn pairwise joint relations with the assumption of *virtual* motion connections among all joints [7, 8, 14, 61], while they rarely exploit key body joints or local body parts that are highly related to walking patterns (*e.g.*, gait [37]) to focus on more discriminative features. Therefore, we can devise *graph motifs* [160], *i.e.*, endowing body-joint nodes with different relational semantic roles based on prior knowledge (*e.g.*, arm-leg collaboration [4]), so as to simultaneously focus on more salient relations such as hierarchical structural relations and gait collaborative relations from skeleton graphs to enhance the skeleton pattern learning for person re-ID.

8.2.2 Skeleton Sample Augmentation

In many skeleton-based tasks, improving the diversity (*e.g.*, multiple views) and amount of training skeleton samples can potentially enhance the model performance [161]. For this purpose, a useful strategy is randomly masking original sequences and their sequential dynamics. This can introduce random perturbation to the model training, which is shown to benefit the model robustness and performance [5]. Firstly, it is possible to achieve multi-granular masking instead of only temporal skeleton masking (*i.e.*, completely masking skeleton frames). For example, similar to Fig. 4.2, we can not only spatially mask the body joints in a skeleton, but also randomly mask the temporal trajectory (*e.g.*, , masking several time steps) of a joint. Using such multi-granular masking for reconstruction or prediction tasks can force the model to infer the structural and temporal context, so as to learn more valuable features and high-level semantics. Secondly, to achieve a more scalable control of randomness, we can explore probabilistic spatial masking to probabilistically and independently mask skeletal structural locations based on a controllable probability distribution (*e.g.*, Bernoulli random variable), and combines probabilistic temporal masking in the same way to generate random partial skeletal motion trajectories, so as to generate more potential spatial-temporal skeleton representations for both semantics learning tasks or model training. Moreover, It is also feasible to apply different augmentation strategies (skeleton rotation, shear, reverse, *etc*) [161] to generate more diverse skeleton sequences, which help better learn the inherent consistency and/or motion continuity of walking and pose patterns.

8.2.3 Importance-Aware Intra-Sequence Learning

While the skeleton prototype learning focuses on relations between different skeleton sequences, the intra-sequence learning aims at capturing inherent relations between different skeletons within a sequence, which can facilitate the model to better select and aggregate effective skeleton features for downstream tasks.

Inspired by the fact that the continuity of human motion typically results in very little variation of poses/skeletons within a small temporal interval [4], a moderate random masking of skeletons in the same sequence can still retain its intra-sequence pattern invariance but introduce random perturbation to help the model enhance

the capture of inherent motion consistency between homologous subsequences. This could potentially encourage the model to learn more effective skeleton representations [5]. In our future work, we will explore more augmentation strategies to generate homologous samples. For example, we can adopt multi-granular masking (as illustrated in Sec. 8.2.2) to partially retain spatial or temporal context of the original sequence, which may facilitate learning more key relations and patterns within the subsequences.

The inference of skeleton importance with the same sequence is also a crucial aspect in intra-sequence relation learning. In the proposed Hi-MPC (see Chapter 5), we have explored the importance of each skeleton within the same sequence based on their normalized relations with corresponding prototypes. It is also feasible to explore the importance of different subsequences to capture key sub-patterns of a sequence, or/and infer the importance of different body parts in representing a recognizable pose. We can combine different-level and cross-level (*e.g.*, sub-component) importance inference mechanisms to mine more valuable relations and more informative features from skeleton data.

8.2.4 Skeleton-based Models for Healthcare

With a concise body representation, small data input, low resource/device requirement, and high convenience in collection through unobtrusive and contactless detection [33], 3D skeletons and their corresponding models have the potential to efficiently capture and learn gait, thereby supporting diverse healthcare-related tasks. In this section, we present several promising directions based on the proposed models in this thesis.

8.2.4.1 Parkinson's Disease Detection and Progression Classification

Our skeleton-based models can be potentially applied to Parkinson's Disease (PD) diagnosis and progression classification. Here we provide a simplified pipeline for this application: First, we can use 3D skeleton sequences to construct the pose, motion or gait representations for PD patients and healthy subjects. In this process, we collect sequence samples of different disease progression stages, and try to keep

a balanced sample size (e.g., similar sample numbers of healthy controls, early-stage PD patients, advanced-stage PD patients) for model learning. Then, we can apply different models to learn effective representations as follows:

- With the proposed SimMC framework (see Chapter 4), we can apply the proposed random masking mechanism to help generate more sub-sequences, which can be exploited by our masked intra-sequence contrastive learning (MIC) scheme to encode the inherent similarity of patterns (*e.g.*, motion continuity and consistency) within sequences. It is worth noting that PD patients might exhibit motion distortion, discontinuity, and inconsistency in their abnormal gait, thus MIC could potentially help learn more effective representations for such gait patterns. Next, we leverage the proposed masked prototype contrastive learning (MPC) to cluster masked skeleton sequences of different classes, so as to learn discriminative features for PD diagnosis (PD vs. healthy controls) and progression classification (*e.g.*, early-stage PD patients vs. advanced-stage PD patients).

- With the proposed Hi-MPC approach (see Chapter 5), we can construct hierarchical representations for each skeleton, to fully depict body structure and motion at different levels (*i.e.*, joint-level, component-level, and limb-level representations). Next, we apply the proposed hierarchical skeleton meta-prototype contrastive learning mechanism to cluster and learn the most typical skeleton features belonging to different classes at multiple body levels, so as to combine them as the multi-level discriminative representations (*i.e.*, multi-level skeleton meta-representation (MSMR)) for PD diagnosis (PD vs. healthy controls) and progression classification (*e.g.*, early-stage PD patients vs. advanced-stage PD patients). To facilitate the skeleton representation learning, the proposed hard skeleton mining (HSM) can be utilized to adaptively infer the informative importance of each skeleton within a sequence. The motivation is that there exist some key skeleton frames in skeleton sequences of PD patients, *e.g.*, some skeleton frames that exhibit more distinguishing motion distortion, discontinuity or inconsistency in their walking/-gait patterns, which can be exploited as unique features in classification and are worth more attention. Therefore, HSM could help the model to better focus on key skeletons associated with PD to learn more effective sequence

representations, and this also provides an intuitive importance inference for gait analysis to improve the interpretability of the model.

- With the proposed TranSG paradigm (see Chapter 6), we can model 3D skeletons as graphs, and apply the proposed skeleton graph transformer (SGT) to fully model structural and actional relations between body joints, which can help capture abnormal body structure and pose features associated with PD. Next, we leverage the proposed graph prototype contrastive learning (GPC) to cluster masked skeleton sequences of different classes, so as to learn discriminative features for PD diagnosis (PD vs. healthy controls) and progression classification (*e.g.*, early-stage PD patients vs. advanced-stage PD patients). Meanwhile, the proposed graph structure-trajectory prompted reconstruction (STPR) can be employed to facilitate the spatial-temporal PD pattern learning and high-level skeleton semantics (*e.g.*, pattern consistency). It is worth noting that PD patients might exhibit motion distortion, discontinuity and inconsistency in their abnormal gait, and these patterns could be reflected in both structural features and trajectory dynamics of skeletons. Therefore, STPR could help the model learn more effective structure-trajectory semantics and representations related to PD.

Finally, we may train a neural network (*e.g.*, MLP) based on the learned features to achieve PD classification or utilize the matching scheme (*e.g.*, probe-gallery matching) for PD prediction. Similarly with the above steps, the proposed models can be further applied to neurodegenerative disease (NDD) diagnosis. The key difference is that we need to collect skeleton sequences from healthy controls and patients of different NDDs. As different NDDs typically possess different skeletal motion and gait patterns, the proposed modules (*e.g.*, STPR, HSM) in different models can also facilitate learning effective skeleton representations for NDDs classification. It is feasible to perform a finer-grained classification by involving different progression stages of NDDs, and we may train a neural network based on the learned features to achieve NDDs classification or utilize the matching scheme for NDDs prediction.

8.2.4.2 Diabetes and Health Monitoring

Based on the 3D human skeleton representations, our skeleton-based models can be potentially applied to monitoring of health or a motion-related disease, *e.g.*,

diabetes. Here we take diabetes as an example to provide a simplified pipeline for this application: First, we can construct pose, motion or gait representations based on a 3D skeleton sequence for diabetic patients and healthy subjects. For example, we can use depth sensors such as Microsoft Kinect deployed at hospitals or home to detect and collect 3D skeleton sequences of them, and exploit the raw sequences, various anthropometric, geometric attributes of skeletons as their motion and gait representations. In this process, we can collect sequence samples of different stages of prediabetes and different progression stages of diabetes, so as to help train a better model to sensitively detect the change of diabetic progression. We may keep a balanced sample size (*e.g.*, similar sample numbers of healthy controls, people with prediabetes, early-stage diabetic patients, advanced-stage diabetic patients) for more robust model learning. Then, we apply different models to learn effective representations as follows:

- With the proposed SimMC framework (see Chapter 4), we can apply the random masking mechanism to help generate more sub-sequences, which can be exploited by the MIC scheme to encode the inherent similarity of patterns (*e.g.*, motion continuity and consistency) within sequences. As we know, diabetic peripheral neuropathy affects patients' motor, sensory, and autonomic nerves, leading to gait abnormalities (*e.g.*, motion distortion, discontinuity or inconsistency), characterized by decreased balance ability and altered plantar pressure. Therefore, MIC could potentially help learn more effective representations for such gait patterns. Next, we leverage the proposed MPC to cluster masked skeleton sequences of different classes, so as to learn discriminative features for diabetes detection (diabetic patients vs. people with prediabetes vs. healthy controls) and progression classification (*e.g.*, early-stage diabetic patients vs. advanced-stage diabetic patients).

- With the proposed Hi-MPC approach (see Chapter 5), we construct hierarchical representations for each skeleton, to fully depict body structure and motion at different levels. Next, we apply the proposed hierarchical skeleton meta-prototype contrastive learning mechanism to cluster and learn the most typical skeleton features belonging to different classes at multiple body levels, so as to combine them as the multi-level discriminative representations (*i.e.*, MSMR) for diabetes detection (diabetic patients vs. people with prediabetes vs. healthy controls) and progression classification (*e.g.*, early-stage

diabetic patients vs. advanced-stage diabetic patients). To facilitate the skeleton representation learning, the proposed HSM mechanism can be utilized to adaptively infer the informative importance of each skeleton within a sequence. As we know, diabetic peripheral neuropathy could lead to gait abnormalities (*e.g.*, motion discontinuity), and they can be characterized by decreased balance ability, altered plantar pressure, abnormal walking speed, unnatural body poses, etc. In this context, there exist some key skeleton frames in skeleton sequences of diabetic patients, *e.g.*, some skeleton frames that exhibit more distinguishing motion distortion, discontinuity or inconsistency in their gait patterns, which can be exploited as unique features in classification and are worth more attention. Therefore, HSM could help the model to better focus on key skeletons associated with diabetes or prediabetes to learn more effective sequence representations, and this also provides an intuitive importance inference for gait analysis to improve the interpretability of the model.

- With the proposed TranSG paradigm (see Chapter 6), we can model 3D skeletons as graphs, and apply the proposed SGT to fully model structural and actional relations between body joints, which can help capture abnormal body structure and pose features associated with diabetes or prediabetes. Next, we leverage the proposed GPC to cluster masked skeleton sequences of different classes, so as to learn discriminative features for diabetes detection (diabetic patients vs. people with prediabetes vs. healthy controls) and progression classification (*e.g.*, early-stage diabetic patients vs. advanced-stage diabetic patients). Meanwhile, the proposed STPR can be employed to facilitate the spatial-temporal PD pattern learning and high-level skeleton semantics (*e.g.*, pattern consistency). As the gait abnormalities (*e.g.*, motion distortion, discontinuity or inconsistency) caused by diabetic peripheral neuropathy may reduce the balance ability of patients (*e.g.*, cause abnormal poses and joint trajectory), STPR could help the model learn more effective structure-trajectory semantics and representations related to diabetes.

Finally, we may train a neural network (*e.g.*, MLP) based on the learned features to achieve such classification or utilize the matching scheme (*e.g.*, probe-gallery matching) for diabetic prediction. By using Kinect to daily collect skeleton sequences from a patient, we can effectively and conveniently monitor the disease

status (*e.g.*, health or different stages of prediabetes or diabetes) using the pre-trained model.

8.2.5 Other Potential Directions and Discussions

- **Skeleton Semantics Learning:** In SimMC (see Chapter 4) and TranSG (see Chapter 6), we have explored masked intra-sequence contrastive learning (MIC) and structure-trajectory prompted reconstruction (STPR) as two skeleton semantics learning tasks to enhance skeleton representation learning. In our future works, we will Incorporate more diverse self-supervised pretext tasks (*e.g.*, prediction, sorting) or unsupervised generative tasks (*e.g.*, pose sequence generation) into prototype contrastive learning to encourage capturing more valuable skeleton semantics for downstream tasks. Moreover, it is a promising direction to explore cross-modal semantics fusion (*e.g.*, language-based gait descriptions) to facilitate gait semantics learning.

- **Generalizable Skeleton Representation:** In Hi-MPC (see Chapter 5), we have devised hierarchical skeleton representations at different levels, which can be slightly modified to be generalized to different datasets to enhance the skeletal body and model modeling. We will explore more unified skeleton representation for different domains (*e.g.*, different scenarios and tasks) to learn generalizable skeleton features, so as to enable the model to be directly applied to multiple domains.

- **Skeleton Sequence Modeling:** In this thesis, we have applied different deep learning paradigms such as MLP, LTSM, Transformer for modeling spatial features and temporal dynamics of skeleton data. In future works, we will explore more efficient and advanced spatial-temporal sequence modeling approaches. For example, the emerging selective state models (*e.g.*, Mamba [162]) can hopefully serve as a general backbone to encode the latent states such as gait states of skeleton data for person re-ID. Moreover, it is possible to introduce more advanced architectures such as Kolmogorov–Arnold Networks (KAN) [163] to enhance the interpretability of skeleton learning process.

- **Multi/Cross-Modal Learning:** Although the focus of our works is to use only 3D skeleton data to perform person re-ID tasks, combining skeleton data

and other modalities such as text (*e.g.*, language-based gait descriptions), RGB images, depth images, radio frequency waves (*e.g.*, Radar waves) is a promising direction, as they can provide pose or gait information from different dimensions (*e.g.*, semantics, appearances, silhouettes) to better identify different persons. Another direction is to transfer and fuse gait representations across skeleton modality and other modality, which can be the key to achieving more general and scalable skeleton learning for more multi-modal tasks.

- **Prompt-Based Skeletal Foundation Model:** Despite the generality of the proposed approaches, they are typically designed for a single modality (*i.e.*, skeleton data) and specific application scenarios such as end-to-end person re-ID tasks, while they may lack user-friendly human-computer interaction to support other downstream tasks. Motivated by the success of Large Language Models (LLMs) and pose generative models [164], we can train or fine-tune them with large-scale skeleton data to build a skeletal foundation model that supports using prompts (*i.e.*, textual user interaction) for skeleton attribute generation (*e.g.*, analyzing and summarizing gait attributes), skeleton augmentation, prediction, classification, and customizable applications (*e.g.*, gait visualization, person re-identification). Constructing this foundational model is advantageous for investigating the scope of adaptability, universality, and interpretability in 3D skeleton data and skeleton-based person re-ID.

- **Unified Evaluation Protocol:** As existing skeleton-based person re-ID studies adopt either direct identity classification [4] or re-ID matching protocol [133] (see Chapter 5), a more comprehensive unified evaluation protocol that provides not only accuracy-related metrics (*e.g.*, Rank-1 accuracy, mAP) but also measures of model generality, robustness, and reliability should be devised. It is also imperative to formulate a fair cross-modality evaluation and comparison protocol that standardizes re-ID settings (*e.g.*, probe/gallery settings, single/multi-shot recognition) for comparing skeleton-based methods and RGB/depth-based methods or multi-modal methods.

- **Model Interpretability:** Improving interpretability of the models is one of the most important focuses in our works. Our first work SimMC can be theoretically modeled as expectation-maximization (EM) solutions (see Chapter

4). We have proved its effectiveness and convergence, which could provide valuable insights for its future improvement and application. In our second work Hi-MPC, we devise the hard skeleton mining mechanism (see Chapter 5), which can help intuitively visualize the most important skeletons and be potentially applied to more healthcare tasks such as Parkinson's disease detection and health monitoring (detailed in Sec. 8.2.4). In our future works, we will design mechanisms to disentangle domain-general and domain-specific skeleton features, which could hopefully help transfer the models to more areas such as gait recognition. For the high-stake fields such as medicine, we will develop more efficient visualization and explanation mechanisms to better verify the importance and effectiveness of skeletons in the decisions of classification or diagnosis.

- **Privacy Protection:** Although existing skeleton-based person re-ID models do not utilize or disclose human appearance information, and all publicly-available training skeleton data are completely anonymized, the privacy issue should be kept in mind when developing this emerging technology further (*e.g.*, combine with RGB images) [165]. As illegally or irresponsibly deploying person re-ID technologies might invade personal privacy, it is still important to establish skeleton-based person re-ID related laws to protect the privacy.

  – **Ethical Statements.** The datasets used in our work are officially shared by reliable research agencies, which guarantee that the collecting, processing, releasing, and using of data have gained the formal consent of participants. To protect privacy, all individuals are anonymized with simple identity numbers. Our models and codes must only be used for legitimate research.

8.3 Impacts and Potential Applications

1. **Industry:**

   - **Security Products:** Skeleton-based person re-ID models can help build light-weight human tracking and identification systems for depth sensor products and companies such as Microsoft Azure Kinect, Intel

RealSense, and Orbbec. There will be a promising market for developing skeleton-specific Integrated Circuit (IC) chips for diverse AI security applications (*e.g.*, integrating 3D pose detection, multi-person identification, and key action prediction) in these products.

- **Mobile Products:** As our models typically require smaller data inputs (*e.g.*, unlabeled skeleton data) and less computational resources than conventional RGB-based methods, they can be integrated into different mobile devices, robots, autopilots, and portable RGB-D devices such as Intel Realsense [32] and Apple Vision Pro to perform efficient identity-related pattern recognition tasks. This will accelerate the development of 3D spatial intelligence of these products. Moreover, our models can be utilized to enhance multi-modal recognition. For example, they can be combined with different sensor data (*e.g.*, camera images, speech data) to jointly perform identity-aware action recognition, gait recognition, motion prediction, and gesture recognition tasks.

2. **Society:**

   - **Smart and Green City:** The skeleton-tracking sensors and skeleton-based models can be potentially deployed into large-scale public camera networks (*e.g.*, in public transport) to enhance smart surveillance and action recognition with 3D skeleton data, which enjoys smaller data inputs, storage, and model sizes. This can facilitate green digital transformation of cities with less energy cost. Moreover, the application of skeleton-based models also brings better privacy protection for the public, as they do not require any visual appearance for recognition.
   - **Public Authentication and Safety:** It is promising to develop skeleton-based ID models and combine them with other biometric features such as faces and fingerprints to enhance public authentication systems and intelligent video surveillance systems. On the other hand, with good robustness to appearance changes and environmental variations [30], our skeleton-based models and frameworks can help track person-of-interest (*e.g.*, criminals) and monitor their pose-based activities under varying scenarios.

3. **Interdisciplinary AI Research (AI + X):**

- **AI + Healthcare:** We can explore disease-related pose/motion/gait classification models based on 3D skeleton patterns. Since many diseases are inherently correlated with abnormal poses, motions or gaits, the pre-trained skeleton-based pose/motion/gait encoding models for person re-ID can be transferred to help classify different patterns in diseases. Furthermore, by using skeleton data captured from non-intrusive depth sensors deployed at home or hospitals, our models can help identify and track a specific person for health monitoring. For example, it can help recognize and record activities (*e.g.*, sitting, walking) to analyze health status. This can be combined with other sensors, such as blood pressure and heart rate sensors, to detect health anomalies such as 3-Highs and accidents such as falls.
- **AI + Gait Analysis:** The skeleton-based pose/gait encoding models for person re-ID can be potentially utilized for different aspects of gait analysis, including but not limited to:
  - Gait Simulation: We can exploit existing skeleton-based gait encoding models, musculoskeletal/anatomical models, and pathological gait models to construct a general gait simulation system for inter-disciplinary research [166].
  - Gait Detection, Classification, and Assessment: It is feasible to leverage 3D skeleton based models to perform abnormal gait (*e.g.*, ataxic gait) detection, disease-related gait (*e.g.*, Parkinsonian gait) classification, and gait-based psychiatric (*e.g.*, depression) assessment, *etc.*
  - Gait Recognition: Gait-based identity recognition based on 3D skeletons can not only be exploited to automate real-time medical analysis (*e.g.*, gait analysis) but also enables concurrently performing other role-based tasks such as skeleton-based action recognition, gesture recognition, interaction recognition, *etc.*
- **AI + Psychology/Neurology:** It is feasible to explore skeleton/pose-based identity-aware psychology (*i.e.*, behavior and emotion) prediction models. In particular, our models can be exploited to learn unique skeleton/pose or gait representations to characterize an individual's different behaviors, and build a mapping between behaviors and emotions [167]

to achieve identity-based emotion (*e.g.*, depression) prediction. Moreover, it is hopeful to construct AI-assisted neurodegenerative disease diagnosis systems based on skeletal gait. Existing studies [33] have demonstrated that human gait features can be a crucial indicator for diagnosing various neurodegenerative diseases (NDDs), which not only provide empirical frameworks for AI-empowered NDDs detection but also envision the application of 3D skeleton data in this area. In this context, our studies on 3D skeleton based human and gait representations could open a new avenue for NDDs digital biomarker modeling [168]. In future, we can apply the AI expertise, domain knowledge of 3D skeletons, gait, and NDDs towards addressing the pressing global challenge of AI-empowered digital biomarker based and contactless early NDDs detection [169]. We believe that this research could not only provide profound benefits to tens of millions of NDDs patients, healthcare professionals, and fellow researchers, but also possess the potential to revolutionize the NDDs related global healthcare market by markedly decreasing the usage of medical resources and cutting annual expenses by billions (US$130 billion per year estimated for global NDDs diagnosis and treatment). Likewise, as discussed in Sec. 8.2.4, our research can also be potentially applied to monitoring other motion-related diseases, such as diabetic patients who possess gait abnormalities (*e.g.*, motion and pose distortion, discontinuity or inconsistency) caused by diabetic peripheral neuropathy. In our future work, we hope to delve deeper into these interdisciplinary directions, and collaborate with medical experts to develop practical clinical systems.

List of Author's Publications During PhD

Conference Proceedings

- **Haocong Rao** and Chunyan Miao. "TranSG: Transformer-Based Skeleton Graph Prototype Contrastive Learning with Structure-Trajectory Prompted Reconstruction for Person Re-Identification," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (**CVPR**)*, pp. 22118-22128, 2023.

- **Haocong Rao** and Chunyan Miao. "SimMC: Simple Masked Contrastive Learning of Skeleton Representations for Unsupervised Person Re-Identification," In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (**IJCAI**)*, pp. 1290-1297, 2022.

- **Haocong Rao** and Chunyan Miao, "Motif Guided Graph Transformer with Combinatorial Skeleton Prototype Learning for Skeleton-Based Person Re-Identification," *Proceedings of the AAAI Conference on Artificial Intelligence (**AAAI**)*, pp. 6703-6712, 2025.

- **Haocong Rao**, Cyril Leung, and Chunyan Miao. "Can ChatGPT Assess Human Personalities? A General Evaluation Framework," In *Findings of Empirical Methods in Natural Language Processing (**EMNLP**)*, pp. 1184–1194, 2023.

Journal Articles[1]

- **Haocong Rao**, Cyril Leung, and Chunyan Miao, "Hierarchical Skeleton Meta-Prototype Contrastive Learning with Hard Skeleton Mining for Unsupervised Person Re-Identification," *International Journal of Computer Vision (**IJCV**)*, vol. 132, pp. 1–23, 2023.

- **Haocong Rao**, Yuan Li, and Chunyan Miao, "Revisiting k-Reciprocal Distance Re-ranking for Skeleton-Based Person Re-Identification," *IEEE Signal Processing Letters (**SPL**)*, vol. 29, pp. 2103–2107, 2022.

- **Haocong Rao***, Minling Zeng*, Xuejiao Zhao, and Chunyan Miao, "A Survey of Artificial Intelligence in Gait-Based Neurodegenerative Disease Diagnosis," *Neurocomputing*, vol. 626, pp. 129533, 2025.

Submitted Papers and Preprints

- **Haocong Rao** and Chunyan Miao. "Recognizing Identities From Human Skeletons: A Survey on 3D Skeleton Based Person Re-Identification," *arXiv preprint arXiv:2401.15296, 2025.*

- **Haocong Rao** and Chunyan Miao, "Deformable Locality-Coordination Graph Motifs for 3D Skeleton Based Person Re-Identification".

- **Haocong Rao** and Chunyan Miao, "General Skeleton Semantics Learning with Probabilistic Masked Context Reconstruction for Skeleton-Based Person Re-Identification".

- **Haocong Rao** and Chunyan Miao. "Skeleton Prototype Contrastive Learning with Multi-Level Graph Relation Modeling for Unsupervised Person Re-Identification," *arXiv preprint arXiv:2208.11814, 2022.*

[1]The superscript * indicates joint first authors

Bibliography

[1] Matteo Munaro, Andrea Fossati, Alberto Basso, Emanuele Menegatti, and Luc Van Gool. One-shot person re-identification with a consumer depth camera. In *Person Re-Identification*, pages 161–181. Springer, 2014. xx, 12, 14, 16, 36, 41, 42, 76, 80, 81, 83, 120, 125, 126, 150

[2] Pietro Pala, Lorenzo Seidenari, Stefano Berretti, and Alberto Del Bimbo. Enhanced skeleton and face 3D data for person re-identification from depth cameras. *Computers & Graphics*, 79:69–80, 2019. xx, 3, 12, 13, 14, 27, 41, 42, 64, 80, 81, 83, 109, 125, 126, 143

[3] Haocong Rao, Siqi Wang, Xiping Hu, Mingkui Tan, Huang Da, Jun Cheng, and Bin Hu. Self-supervised gait encoding with locality-aware attention for person re-identification. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 1, pages 898–905, 2020. xx, 3, 4, 5, 13, 14, 19, 20, 27, 41, 42, 44, 45, 64, 65, 81, 82, 85, 90, 91, 125, 126, 127, 130

[4] Haocong Rao, Siqi Wang, Xiping Hu, Mingkui Tan, Yi Guo, Jun Cheng, Xinwang Liu, and Bin Hu. A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6649–6666, 2021. xx, 3, 5, 13, 17, 19, 20, 33, 39, 41, 42, 44, 65, 77, 78, 80, 81, 82, 83, 85, 90, 91, 106, 109, 111, 117, 122, 124, 125, 126, 127, 130, 143, 151, 152, 159, 160, 167

[5] Haocong Rao and Chunyan Miao. SimMC: Simple masked contrastive learning of skeleton representations for unsupervised person re-identification. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1290–1297, 2022. xx, 2, 5, 12, 19, 27, 56, 57, 64, 65, 69, 70, 76, 77, 78, 80, 81, 82, 83, 85, 88, 90, 91, 93, 101, 103, 106, 109, 116, 120, 122, 124, 125, 126, 127, 129, 130, 136, 143, 144, 145, 150, 151, 152, 160, 161

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, volume 96, pages 226–231, 1996. xx, 32, 57, 70, 78, 103, 124, 129

[7] Haocong Rao, Shihao Xu, Xiping Hu, Jun Cheng, and Bin Hu. Multi-level graph encoding with structural-collaborative relation learning for skeleton-based person re-identification. In *International Joint Conference on Artificial*

Intelligence (IJCAI), pages 973–980, 2021. xxi, 3, 4, 13, 14, 15, 19, 20, 24, 25, 27, 39, 41, 42, 68, 77, 80, 81, 83, 87, 90, 109, 110, 122, 125, 130, 151, 152, 159

[8] Haocong Rao, Xiping Hu, Jun Cheng, and Bin Hu. SM-SGE: A self-supervised multi-scale skeleton graph encoding framework for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1812–1820, 2021. xxi, 3, 4, 14, 15, 19, 24, 25, 39, 41, 42, 44, 45, 64, 77, 81, 82, 83, 85, 87, 88, 90, 91, 113, 119, 122, 123, 125, 127, 128, 130, 151, 152, 159

[9] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(2):207–244, 2009. xxi, 44, 85, 127, 128

[10] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *International Conference on Machine Learning (ICML)*, pages 209–216, 2007. xxi, 44, 85, 127

[11] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 262–275. Springer, 2008. xxi, 44, 85, 127

[12] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2360–2367. IEEE, 2010. xxi, 11, 44, 85, 127

[13] Zheng Liu, Zhaoxiang Zhang, Qiang Wu, and Yunhong Wang. Enhancing person re-identification by integrating gait biometric. *Neurocomputing*, 168: 1144–1156, 2015. xxi, 37, 38, 44, 77, 85, 120, 121, 127, 128

[14] Haocong Rao and Chunyan Miao. TranSG: Transformer-based skeleton graph prototype contrastive learning with structure-trajectory prompted reconstruction for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22118–22128, 2023. xxii, 19, 109, 117, 159

[15] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[16] Athira Nambiar, Alexandre Bernardino, and Jacinto C Nascimento. Gait-based person re-identification: A survey. *ACM Computing Surveys*, 52(2): 33, 2019. 1

[17] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):591–606, 2015.

[18] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. Sarc3D: a new 3D body model for people tracking and re-identification. In *International Conference on Image Analysis and Processing*, pages 197–206. Springer, 2011.

[19] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys*, 46(2):29, 2013.

[20] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry Steven Davis, and Wen Gao. Multi-task learning with low rank attribute embedding for multi-camera person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1167–1181, 2018. 1

[21] Global digital identity market report. `http://bit.ly/digital-identity-market-report`, August 2022. 1

[22] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark J Finocchio, Richard Moore, Alex Abenathar Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1297–1304, 2011. 1, 38, 121, 143

[23] Matteo Munaro, Alberto Basso, Andrea Fossati, Luc Van Gool, and Emanuele Menegatti. 3D reconstruction of freely moving persons for re-identification with a depth sensor. In *International Conference on Robotics and Automation (ICRA)*, pages 4512–4519. IEEE, 2014. 1, 3, 13, 16, 27, 64

[24] Virginia O Andersson and Ricardo M Araujo. Person identification using anthropometric and gait data from Kinect sensor. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 425–431, 2015. 3, 12, 14, 27, 36, 64, 76, 109, 120, 143

[25] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020. 3, 4, 13, 14, 15, 27, 37, 38, 41, 42, 64, 76, 80, 81, 83, 90, 91, 109, 121, 125, 126, 128, 130, 143

[26] Haocong Rao and Chunyan Miao. A survey on 3D skeleton based person re-identification: Approaches, designs, challenges, and future directions. *arXiv preprint arXiv:2401.15296*, 2024. 1

[27] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518, 2003. 2

[28] Chen Wang, Junping Zhang, Liang Wang, Jian Pu, and Xiaoru Yuan. Human identification using temporal information preserving gait template. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2164–2176, 2011.

[29] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object Re-ID. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 11309–11321, 2020. 2, 17, 18, 39, 78

[30] Fei Han, Brian Reily, William Hoff, and Hao Zhang. Space-time representation of people based on 3D skeletal data: A review. *Computer Vision and Image Understanding*, 158:85–105, 2017. 2, 12, 169

[31] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2021. 2

[32] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–10, 2017. 2, 169

[33] Haocong Rao, Minlin Zeng, Xuejiao Zhao, and Chunyan Miao. A survey of artificial intelligence in gait-based neurodegenerative disease diagnosis. *arXiv preprint arXiv:2405.13082*, 2024. 2, 161, 171

[34] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with RGB-D sensors. In *European Conference on Computer Vision (ECCV) Workshop*, pages 433–442. Springer, 2012. 3, 12, 13, 27, 64

[35] Jang-Hee Yoo, Mark S Nixon, and Chris J Harris. Extracting gait signatures based on anatomical knowledge. In *Proceedings of BMVA Symposium on Advancing Biometric Technologies*, pages 596–606. Citeseer, 2002. 3, 12

[36] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3, 20, 109

[37] M Pat Murray, A Bernard Drought, and Ross C Kory. Walking patterns of normal men. *Journal of Bone and Joint Surgery*, 46(2):335–360, 1964. 4, 5, 31, 69, 113, 159

[38] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019. 4

[39] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5, 18, 19, 66, 72

[40] Shihao Xu, Haocong Rao, Xiping Hu, Jun Cheng, and Bin Hu. Prototypical contrast and reverse prediction: Unsupervised skeleton based action recognition. *IEEE Transactions on Multimedia*, 25:624–634, 2021. 5, 19

[41] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 1–16. Springer, 2014. 11

[42] Zhong Zhang, Haijia Zhang, and Shuang Liu. Person re-identification using heterogeneous local graph attention networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12136–12145, 2021. 11

[43] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 391–401. Springer, 2012. 11

[44] Rahul Rama Varior, Gang Wang, Jiwen Lu, and Ting Liu. Learning invariant color features for person reidentification. *IEEE Transactions on Image Processing*, 25(7):3395–3410, 2016. 11

[45] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206, 2015. 11

[46] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 23(8):3656–3670, 2014.

[47] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2012. 11

[48] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 152–159, 2014. 11

[49] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. 11, 40, 41, 80, 124, 125, 147, 151

[50] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 79–88, 2018. 11

[51] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018. 11

[52] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2138–2147, 2019. 11

[53] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. ABD-Net: Attentive but diverse person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8351–8361, 2019. 11

[54] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 667–676, 2019. 11

[55] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–18, 2018. 12

[56] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 33, pages 8738–8745, 2019. 12

[57] Jinlin Wu, Yang Yang, Hao Liu, Shengcai Liao, Zhen Lei, and Stan Z Li. Unsupervised graph association for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8321–8330, 2019. 12

[58] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 994–1003, 2018. 12

[59] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 598–607, 2019. 12

[60] Yuhang Ding, Hehe Fan, Mingliang Xu, and Yi Yang. Adaptive exploration for unsupervised person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(1):1–19, 2020. 12

[61] Haocong Rao and Chunyan Miao. Skeleton prototype contrastive learning with multi-level graph relation modeling for unsupervised person re-identification. *arXiv preprint arXiv:2208.11814*, 2022. 15, 24, 109, 110, 116, 124, 125, 126, 127, 129, 130, 136, 138, 159

[62] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3960–3969, 2017. 15

[63] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. GLAD: Global-local-alignment descriptor for pedestrian retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 420–428, 2017. 15

[64] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pages 2540–2549, 2022. 16

[65] Jiaxuan Lu, Hai Wan, Peiyan Li, Xibin Zhao, Nan Ma, and Yue Gao. Exploring high-order spatio–temporal correlations from skeleton for person re-identification. *IEEE Transactions on Image Processing*, 32:949–963, 2023. 16

[66] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4099–4108, 2018. 16

[67] Sabesan Sivapalan, Daniel Chen, Simon Denman, Sridha Sridharan, and Clinton Fookes. Gait energy volumes and frontal gait recognition using depth images. In *International Joint Conference on Biometrics*, pages 1–6. IEEE, 2011. 16

[68] Lin Chunli and Wang Kejun. A behavior classification based on enhanced gait energy image. In *International Conference on Networking and Digital Society*, volume 2, pages 589–592. IEEE, 2010. 16

[69] Albert Haque, Alexandre Alahi, and Li Fei-Fei. Recurrent attention models for depth-based person identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1229–1238, 2016. 16, 20, 109

[70] Mohamed Hasan and Noborou Babaguchi. Long-term people reidentification using anthropometric signature. In *International Conference on Biometrics Theory, Applications and Systems*, pages 1–6. IEEE, 2016. 16

[71] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Robust depth-based person re-identification. *IEEE Transactions on Image Processing*, 26(6):2588–2603, 2017. 16

[72] Ziyang Wang, Dan Wei, Xiaoqiang Hu, and Yiping Luo. Human skeleton mutual learning for person re-identification. *Neurocomputing*, 388:309–323, 2020. 16

[73] Vuong D Nguyen, Samiha Mirza, Pranav Mantini, and Shishir K Shah. Attention-based shape and gait representations learning for video-based cloth-changing person re-identification. *arXiv preprint arXiv:2402.03716*, 2024. 16

[74] Nikolaos Karianakis, Zicheng Liu, Yinpeng Chen, and Stefano Soatto. Reinforced temporal attention and split-rate transfer for depth-based person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 715–733. Springer, 2018. 16

[75] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1970–1979, 2017. 16

[76] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *Proceedings of the European conference on computer vision (ECCV)*, pages 54–70, 2018. 16

[77] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5380–5389, 2017. 16

[78] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3623–3632, 2019. 16

[79] Xiang Li, Wei-Shi Zheng, Xiaojuan Wang, Tao Xiang, and Shaogang Gong. Multi-scale learning for low-resolution person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3765–3773, 2015. 16

[80] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8042–8051, 2018. 16

[81] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742, 2018. 17, 103

[82] Taihong Xiao, Sifei Liu, Shalini De Mello, Zhiding Yu, Jan Kautz, and Ming-Hsuan Yang. Learning contrastive representation for semantic correspondence. *International Journal of Computer Vision*, 130(5):1293–1309, 2022.

[83] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 26

[84] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, pages 4904–4916. PMLR, 2021. 17, 26

[85] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 11834–11845, 2021. 17

[86] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, pages 9929–9939, 2020. 17, 50, 58, 103

[87] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9912–9924, 2020. 17

[88] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representation (ICLR)*, 2021. 18, 56, 57, 64, 69, 101, 103, 106

[89] Xudong Wang, Ziwei Liu, and Stella X Yu. Unsupervised feature learning by cross-level instance-group discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12586–12595, 2021. 17, 18

[90] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010. 17

[91] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020. 17, 57, 62, 102, 103, 106

[92] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 22243–22255, 2020. 17

[93] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 17

[94] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 17, 57, 102, 103, 106

[95] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021. 17, 33, 34, 48, 61, 62

[96] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations (ICLR)*, 2021. 18

[97] Sangryul Jeon, Dongbo Min, Seungryong Kim, and Kwanghoon Sohn. Mining better samples for contrastive learning of temporal correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1034–1044, 2021. 18

[98] Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1074–1083, 2021. 18, 19

[99] Weilun Wang, Wengang Zhou, Jianmin Bao, Dong Chen, and Houqiang Li. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 14020–14029, 2021. 18, 19

[100] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances*

in Neural Information Processing Systems (NeurIPS), 33:21798–21809, 2020. 18, 19

[101] Vikas Verma, Thang Luong, Kenji Kawaguchi, Hieu Pham, and Quoc Le. Towards domain-agnostic contrastive learning. In *International Conference on Machine Learning (ICML)*, pages 10530–10541. PMLR, 2021.

[102] Shaofeng Zhang, Meng Liu, Junchi Yan, Hengrui Zhang, Lingxiao Huang, Xiaokang Yang, and Pinyan Lu. M-mix: Generating hard negatives via multi-sample mixing for contrastive learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2461–2470, 2022. 18, 19

[103] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 19

[104] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-RNN: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5308–5317, 2016. 19

[105] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1227–1236, 2019.

[106] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1112–1121, 2020.

[107] Tianjiao Li, Qiuhong Ke, Hossein Rahmani, Rui En Ho, Henghui Ding, and Jun Liu. Else-Net: Elastic semantic network for continual action recognition from skeleton data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13434–13443, 2021. 19

[108] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 22, 71, 115

[109] Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5884–5888. IEEE, 2018. 22

[110] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[111] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[112] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in neural information processing systems*, 34:15908–15919, 2021. 22

[113] David A Winter. *Biomechanics and motor control of human movement*. John Wiley & Sons, 2009. 24, 68

[114] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 214–223, 2020. 24, 25, 119, 135

[115] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2463–2473, 2019. 26

[116] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 26

[117] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 26

[118] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, 2020. 26

[119] Athira Nambiar, Alexandre Bernardino, Jacinto C Nascimento, and Ana Fred. Context-aware person re-identification in the wild via fusion of gait and anthropometric features. In *International Conference on Automatic Face & Gesture Recognition*, pages 973–980. IEEE, 2017. 35, 76, 120, 150

[120] Matteo Munaro, Stefano Ghidoni, Deniz Tartaro Dizmen, and Emanuele Menegatti. A feature-based approach to people re-identification using skeleton keypoints. In *International Conference on Robotics and Automation (ICRA)*, pages 5644–5651. IEEE, 2014. 36, 76, 120, 150

[121] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *International Conference on Pattern Recognition (ICPR)*, volume 4, pages 441–444. IEEE, 2006. 37, 38, 76, 120, 121

[122] Ching-Hang Chen and Deva Ramanan. 3D human pose estimation= 2D pose estimation+ matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7035–7043, 2017. 38, 85, 121, 122

[123] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2019. 38, 85, 121, 122

[124] Rui Zhao, Kang Wang, Hui Su, and Qiang Ji. Bayesian graph convolution lstm for skeleton based action recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6882–6892, 2019. 39, 77, 122, 151

[125] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. *arXiv preprint arXiv:2103.11568*, 2021. 39, 78

[126] Lutz Prechelt. Early stopping-but when? In *Advances in Neural Information Processing Systems (NeurIPS) Workshop*, pages 55–69, 1998. 40, 79, 124

[127] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1318–1327, 2017. 40, 80, 124, 143, 144, 147, 148, 149, 151, 152

[128] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008. 45, 88, 129

[129] Brian C Ross. Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357, 2014. 50, 96

[130] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910, 2021. 50, 58, 103

[131] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *arXiv preprint arXiv:2201.08164*, 2022. 57, 102

[132] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 57, 102

[133] Haocong Rao, Cyril Leung, and Chunyan Miao. Hierarchical skeleton meta-prototype contrastive learning with hard skeleton mining for unsupervised person re-identification. *International Journal of Computer Vision*, 132:1–23, 2023. 64, 109, 125, 159, 167

[134] Joey Tianyi Zhou, Sinno Jialin Pan, and Ivor W. Tsang. A deep learning framework for hybrid heterogeneous transfer learning. *Artificial Intelligence*, 275:310–328, 2019. doi: 10.1016/j.artint.2019.06.001. 70

[135] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 30, pages 2058–2065, 2016. 71, 88, 95

[136] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation OSDI 16)*, pages 265–283, 2016. 91, 130

[137] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, pages 535–547, 2019. 91, 130

[138] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010. 96

[139] Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2007. 105, 106

[140] CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983. 105, 106

[141] Victor Hasselblad. Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association*, 64(328): 1459–1471, 1969.

[142] John H Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate behavioral research*, 5(3):329–350, 1970.

[143] Nan Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

[144] Donald B Rubin and Dorothy T Thayer. EM algorithms for ml factor analysis. *Psychometrika*, 47(1):69–76, 1982. 105

[145] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. In *AAAI Conference on Artificial Intelligence (AAAI) Workshop*, 2021. 114, 115, 123, 136

[146] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 115

[147] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning (ICML)*, pages 448–456. PMLR, 2015. 115

[148] Vijay Prakash Dwivedi, Chaitanya K Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 24(43):1–48, 2023. 123, 136

[149] Haocong Rao, Yuan Li, and Chunyan Miao. Revisiting k-reciprocal distance re-ranking for skeleton-based person re-identification. *IEEE Signal Processing Letters*, 29:2103–2107, 2022. 143

[150] Xiaohui Shen, Zhe Lin, Jonathan Brandt, Shai Avidan, and Ying Wu. Object retrieval and localization with spatially-constrained similarity measure and k-NN re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3013–3020. IEEE, 2012. 143

[151] Mang Ye, Chao Liang, Yi Yu, Zheng Wang, Qingming Leng, Chunxia Xiao, Jun Chen, and Ruimin Hu. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia*, 18(12):2553–2566, 2016. 143

[152] Wei Li, Yang Wu, Masayuki Mukunoki, and Michihiko Minoh. Common-near-neighbor analysis for person re-identification. In *IEEE International Conference on Image Processing*, pages 1621–1624. IEEE, 2012. 144

[153] Jorge Garcia, Niki Martinel, Christian Micheloni, and Alfredo Gardel. Person re-identification ranking optimisation by discriminant context information analysis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1305–1313, 2015. 144

[154] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2501–2514, 2016. 144

[155] Rui Zhao, Wanli Oyang, and Xiaogang Wang. Person re-identification by saliency learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):356–370, 2017.

[156] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Unsupervised person re-identification by deep asymmetric metric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):956–973, 2020. 144

[157] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 33, pages 8126–8133, 2019. 144, 146, 147

[158] Danfeng Qin, Stephan Gammeter, Lukas Bossard, Till Quack, and Luc Van Gool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 777–784. IEEE, 2011. 147

[159] Song Bai and Xiang Bai. Sparse contextual activation for efficient visual re-ranking. *IEEE Transactions on Image Processing*, 25(3):1056–1069, 2016. 148

[160] Yu-Hui Wen, Lin Gao, Hongbo Fu, Fang-Lue Zhang, Shihong Xia, and Yong-Jin Liu. Motif-GCNs with local and non-local temporal blocks for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2009–2023, 2022. 159

[161] Haocong Rao, Shihao Xu, Xiping Hu, Jun Cheng, and Bin Hu. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. *Information Sciences*, 569:90–109, 2021. 160

[162] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 166

[163] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024. 166

[164] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. PoseGPT: Quantization-based 3D human motion generation and forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 417–435. Springer, 2022. 167

[165] Mang Ye, Wei Shen, Junwu Zhang, Yao Yang, and Bo Du. Securereid: Privacy-preserving anonymization for person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:2840–2853, 2024. 168

[166] Jungnam Park, Sehee Min, Phil Sik Chang, Jaedong Lee, Moon Seok Park, and Jehee Lee. Generative gaitnet. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 170

[167] Elisabeta Marinoiu, Mihai Zanfir, Vlad Olaru, and Cristian Sminchisescu. 3D human sensing, action and emotion recognition in robot assisted therapy of children with autism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2158–2167, 2018. 170

[168] Yingcheng Liu, Guo Zhang, Christopher G Tarolli, Rumen Hristov, Stella Jensen-Roberts, Emma M Waddell, Taylor L Myers, Meghan E Pawlik, Julia M Soto, Renee M Wilson, et al. Monitoring gait at home with radio waves in Parkinson's disease: A marker of severity, progression, and medication response. *Science Translational Medicine*, 14(663):eadc9669, 2022. 171

[169] Yuzhe Yang, Yuan Yuan, Guo Zhang, Hao Wang, Ying-Cong Chen, Yingcheng Liu, Christopher G Tarolli, Daniel Crepeau, Jan Bukartyk, Mithri R Junna, et al. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nature medicine*, 28(10):2207–2215, 2022. 171