DOMAIN ADAPTATION FOR VIDEO ACTION RECOGNITION

Xiyu Wang

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Master of Engineering

2023

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.

2 May 2023

Date

Xiyu Wang

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 accordance with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

18 May 2023

Date

Mao Kezhi

Prof. Kezhi Mao

Authorship Attribution Statement

This thesis contains material from 1 paper accepted at peer-reviewed conferences in which I am listed as an author.

Chapter 3 is published as Xiyu Wang, Yuecong Xu, Jianfei Yang, and Kezhi Mao. "Calibrating class weights with multi-modal information for partial video domain adaptation." In Proceedings of the 30th ACM International Conference on Multimedia, pp. 3945-3954. 2022.

The contributions of the co-authors are as follows:

- Prof Mao provided the initial research direction and reviewed the final manuscript draft.
- I prepared the manuscript drafts, and Dr. Yuecong And Dr. Jianfei revised them.
- Dr. Yuecong is involved in the development of part of the code and some baseline testing while I conducted most of the development.

2 May 2023

.....

Date

.....

Xiyu Wang

Acknowledgements

I wish to express my gratitude to my advisor, co-authors, and my parents. Prof. Mao gave me some initial thoughts on what I should do as a fresh M.Eng. student, and I appreciate his help in introducing Dr. Yuecong to me. I also want to thank Dr. Yuecong for teaching me a lot about designing and conducting experiments and drafting a comprehensive draft for submission. Those abilities are essential for a researcher. Dr. Jianfei also provided much guidance when I was confused about experiment results as well as my research direction. Finally, as a self-funded M.Eng. student, I want to express my greatest gratitude to my parents, who funded me throughout my M.Eng. candidature. They also provided me with comfort when I was facing difficulties. Overall, I am thankful to people who had helped me along the way and I simply cannot express my gratitude enough for them.

Contents

Abstract

Humans can effortlessly learn from a specific data distribution and generalize well to various situations without excessive supervision. In contrast, deep learning models often struggle to achieve similar generalization capabilities. This is primarily because deep models are trained with algorithms that aim to minimize empirical risks on training data and assume that test data share the same distribution as train data. However, significant domain shifts between training (source) and testing (target) data can occur, causing deep models to generalize poorly on target domains and necessitating additional supervision for adaptation.

To address this, Video-based Unsupervised Domain Adaptation (VUDA) has been proposed as a cost-efficient approach for transferring video action recognition models from the source domain to an unlabeled target domain. Nonetheless, VUDA relies on strong assumptions, such as identical label spaces and fixed target domains, which may not hold true in real-world applications. Consequently, this thesis aims to eliminate these assumptions to broaden the applicability of video adaptation methods, focusing on two major shortcomings of conventional VUDA methods, e.g., partial domain adaptation (adapting from a source domain with many classes to a target domain with fewer classes) and continual domain adaptation (adapting to continuously changing target domains). For partial domain adaptation, this thesis proposes the Multi-modality Cluster-calibrated partial Adversarial Network (MCAN), which constructs a multi-modal network to extract robust features and a novel calibration method to refine target class distribution estimation, effectively filtering out irrelevant source classes. To further address some real challenges in the field of adapting deep video models, the problem of continuous video domain adaptation is defined and this thesis proposes Confidence-Attentive network with geneRalization enhanced self-knowledge disTillation (CART). This method leverages attentive learning and a novel data generalization enhanced self-knowledge distillation to preserve previously learned knowledge on seen target domains while adapting to newly encountered ones, ultimately providing a performative model for multiple seen target domains at a minimal cost.

This thesis evaluates the proposed partial and continuous video domain adaptation methods on existing and newly constructed benchmarks in this thesis. Our results demonstrated significant performance improvements for MCAN and CART, with MCAN showing particularly strong gains when domain shifts were substantial and CART demonstrating a superior capability of preserving learned knowledge. In conclusion, our research findings on partial and continuous domain adaptation effectively broadened the applicability of video domain adaptation methods, making them more general and cost-efficient.

List of Figures

List of Tables

Symbols and Acronyms

Symbols

$\mathcal{H}$	the hypothesis space
$\mathcal{H}\Delta\mathcal{H}$	the symmetric difference hypothesis space with respect to the hypothesis space $\mathcal{H}$
ϵ	the expected error of hypothesis h
$\mathbf{C}_0$	a constant term
$\sum_{i=a}^{b} f(i)$	the sum of function $f(i)$ for variable i with values from a to b
$\|\|\cdot\|\|^2$	L2-norm
δ	softmax function
$KL(\cdot \mid \cdot)$	Kullback-Leibler divergence
$\mathbb{1}(\cdot)$	A choice function, when the condition specified is true, it equals 1, otherwise 0

Acronyms

2D CNN	2-Dimensional Convolution Neural Network
3D CNN	3-Dimensional Convolution Neural Network
ADDA	Adversarial Discriminative Domain Adaptation
BoVW	Bag-of-Visual Words
CADA	Consensus Adversarial Domain Adaptation
CART	geneRalization enhanced self-knowledge disTillation
CiDA	Class-incremental Domain Adaptation
CL	Continual Learning
CMA	Continuous Manifold-based Adaptation
CNN	Convolutional Neural Network
CORAL	CORrelation ALignment

CSN	Channel-Separated convolutional Networks
CDA	Continuous Domain Adaptation
CVDA	Continuous Video Domain Adaptation
CyCADA	Cycle-Consistent Adversarial Domain Adaptation
DA	Domain Adaptation
DANN	Domain Adversarial Neural Network
DT	Dense Trajectory
DLMM	Differentiated Learning for Multi-Modal domain adaptation
ERM	Empirical Risk Minimization
GAN	Generative Adversarial Network
GPU	Graphical Processing Unit
HOG	Histogram of Oriented Gradients
iDT	improved Dense Trajectory
IADA	Incremental Adversarial Domain Adaptation
KD	Knowledge Distillation
MBH	Motion Boundary Histogram
MCAN	Multi-modality Cluster-calibrated partial Adversarial Network
MEI	Motion Energy Image
MHI	Motion History Image
MMD	Maximum Mean Discrepancy
PATAN	Partial Adversarial Temporal Attentive Network
PVDA	Partial Video Domain Adaptation
ResNet	Residual Neural Network
RKHS	Reproducing Kernel Hilbert Space
SAVA	Shuffle and Attend: Video domain Adaptation
SDT	Structured Dense Trajectories
SFDA	Source-Free Domain Adaptation
SIE	Statistically Invariant Sample Selection
SISS	Statistically Invariant Embedding
SIFT	Scale-Invariant Feature Transformation
STIPs	Space-Time Interest Points
SVM	Support Vector Machine
TA^3N	Temporal Attentive Adversarial Adaptation Network
TCoN	Temporal Co-attention Network
TRN	Temporal Relation Network

TSN	Temporal Segment Network
TSM	Temporal Shift Module
UDA	Unsupervised Domain Adaptation
ViT	Vision Transformer
ViViT	Video Vision Transformer
VGG	Visual Geometry Group
VUDA	Video-based Unsupervised Domain Adaptation
i.i.d.	independent and identically distributed

Chapter 1

Introduction

1.1 Overview and Motivation

As one of the most prevalent data types encountered daily, understanding video content through deep learning methods has gained significant interest in recent decades. A fundamental aspect of developing the video understanding capability of AI is vision-based action recognition, which has numerous applications in surveillance, smart homes, autonomous driving, and robotics. While handcrafted features were initially employed in vision-based action recognition models, deep learning methods, such as convolutional neural networks (CNNs), have gradually taken the lead in performance. Pioneering deep learning-based research [2] in vision-based action recognition focused on extracting spatial features from individual frames and then fusing them using a late-fusion approach to obtain spatial-temporal representations. However, it was discovered that capturing temporal information, such as motion, proved challenging for CNNs, as they are optimized for processing geometric relationships. In response to subsequent research [6, 7] expanded 2-dimensional CNNs and proposed 3-dimensional CNNs to concurrently process spatial and temporal information. To date, as self-attention-based transformer networks have gained prominence in vision-based action recognition, a consensus among researchers remains that accurately modeling temporal relations within videos is crucial for enabling precise video-based action recognition.

Modern deep neural networks often overfit training datasets due to their increased network capacity in recent years. Consequently, deep learning methods may underperform on target datasets with different data distributions than the training dataset. This is because the model can reach a local optimum for the training dataset without being optimal for other data distributions. To address this issue, a line of work called Unsupervised Domain Adaptation (UDA) for image-based models was proposed. UDA aims to mitigate the challenges of obtaining costly and unrealistic fine-tuning datasets. The founding theory of UDA suggests that, given a fixed model and datasets, adaptation performance is largely determined by the domain discrepancy between the extracted features of the source and target datasets. Consequently, related research [8, 9] generally aims to minimize such discrepancies using strategies like adversarial domain adaptation or minimizing statistical approximations of domain discrepancies. For videos, adaptation is complicated by the increased complexity of video features due to the additional temporal dimension. Although capturing temporal features is necessary for video action recognition models, current deep learning methods sometimes focus on spatial information and fail to learn temporal relations. This failure can lead to poor domain adaptation performance, prompting researchers to explore Video-based Unsupervised Domain Adaptation (VUDA). Recent VUDA work [10, 11] focuses on aligning temporal features to improve overall domain adaptation performance in video action recognition models. These works typically emphasize independent processing and alignment of temporal features during the adaptation process. While VUDA has achieved promising results on video domain adaptation benchmarks, it has limitations when applied to real-world scenarios. Specifically, VUDA methods generally assume that the label spaces between the source and target domains are identical, which can be impractical since identifying target classes in the target dataset can be exhaustive. Additionally, VUDA methods do not consider cases where the model needs to adapt to continuously arriving target domains, even though collecting video datasets can be time-consuming and samples may arrive in batches.

In the first case, known as Partial Domain Adaptation (PDA), the target label space is assumed to be a subset of the source label space. This assumption introduces the challenge of adapting to the target domain while avoiding negative transfers of outlier classes or classes that exist only in the source dataset. For example, a model may be confused by two similar actions, such as walking and

running. If the source dataset contains both actions while the target dataset only has 'running,' the model may misclassify all 'running' samples in the target dataset as 'walking,' leading to degraded adaptation performance. Existing image-based methods address this issue by estimating the target class distribution based on aggregated target predictions. For Partial Video Domain Adaptation (PVDA), temporal features are critical for model transfer, prompting some methods to extract additional information from temporal relations to improve target class estimation accuracy. This thesis further explores leveraging temporal information for efficient PVDA and proposes a novel strategy to refine estimation based on certain clustering structures. A detailed discussion of these methods is found in Chapter 3.

In the second case, Continuous Video Domain Adaptation (CVDA), a deep model is continuously adapted to new domains without supervision on both the source and target domains. CVDA remains largely unexplored in academia, as it is a relatively new topic. Conventional VUDA methods may struggle in such cases, as they are not designed to adapt to target datasets arriving in separate pieces, each with potentially different data distributions. To address this, the thesis first formulates the new CVDA problem and conducts extended experiments on this topic. Based on insights into how modern video action recognition models respond to CVDA scenarios, the thesis proposes novel training policies and regularization methods to prevent catastrophic forgetting of previously learned knowledge. Chapter 4 provides a detailed explanation of the problem and a comprehensive illustration of the proposed solution.

Overall, this thesis introduces the general concept of Video-based Unsupervised Domain Adaptation (VUDA), identifying its limitations in Partial Video Domain Adaptation (PVDA) and Continuous Video Domain Adaptation (CVDA). Building on VUDA methods, the thesis proposes solutions to these challenges in Chapters 3 and 4, aiming to enable more practical and general domain adaptation for deep video models.

1.2 Major Contributions

One of the major contributions of this thesis is that it enables more efficient and accurate domain adaptation for video action recognition models in more practical scenarios, e.g., partial domain adaptation scenarios and continuous domain adaptation scenarios. More specifically, they can be stated as follows:

Multi-modality Cluster-calibrated partial Adversarial Network (MCAN) Assuming the source label space subsumes the target one, Partial Video Domain Adaptation (PVDA) is a more general and practical scenario for cross-domain video classification problems. The key challenge of PVDA is to mitigate the negative transfer caused by the source-only outlier classes. To tackle this challenge, a crucial step is to aggregate target predictions to assign class weights by up-weighing target classes and down-weighing outlier classes. However, incorrect predictions of class weights can mislead the network and lead to negative transfer. Previous works improve the class weight accuracy by utilizing temporal features and attention mechanisms, but these methods may fall short when trying to generate accurate class weight when domain shifts are significant, as in most real-world scenarios. To deal with these challenges, this thesis first proposes the Multi-modality partial Adversarial Network (MAN), which utilizes multi-scale and multi-modal information to enhance PVDA performance. Based on MAN, this thesis then proposes Multi-modality Cluster-calibrated partial Adversarial Network (MCAN). It utilizes a novel class weight calibration method to alleviate the negative transfer caused by incorrect class weights. Specifically, the calibration method tries to identify and weigh correct and incorrect predictions using distributional information implied by unsupervised clustering. Extensive experiments are conducted on prevailing PVDA benchmarks, and the proposed MCAN achieves significant improvements when compared to state-of-the-art PVDA methods.

Confidence-Attentive network with geneRalization enhanced self-knowledge disTillation (CART) Continuous Video Domain Adaptation (CVDA) is a scenario where a source model is required to adapt to a series of individually available changing target domains continuously without source data or target supervision. It has wide applications, such as robotic vision and autonomous driving. The main underlying challenge of CVDA is to learn helpful information only from the unsupervised target data while avoiding forgetting previously

learned knowledge catastrophically, which is out of the capability of previous Video-based Unsupervised Domain Adaptation methods. Therefore, this thesis proposes a Confidence-Attentive network with geneRalization enhanced self-knowledge disTillation (CART) to address the challenge in CVDA. Firstly, to learn from unsupervised domains, pseudo labels are utilized. However, in continuous adaptation, prediction errors can accumulate rapidly in pseudo labels, and CART effectively tackles this problem with two key modules. Specifically, The first module generates refined pseudo labels using model predictions and deploys a novel attentive learning strategy. The second module compares the outputs of augmented data from the current model to the outputs of weakly augmented data from the source model, forming a novel consistency regularization on the model to alleviate the accumulation of prediction errors. Extensive experiments suggest that the CVDA performance of CART outperforms existing methods by a considerable margin.

1.3 Thesis Organization

Chapter 1 introduces the background and scope of this thesis and provides a brief overview of domain adaptation for video action recognition as well as some practical limits of standard unsupervised video-based domain adaptation. Those practical limits effectively motivated the works contained in this thesis. The main contributions and outline of the thesis are also briefly listed.

Chapter 2 reviews the research progress of prevailing video action recognition methods and domain adaptation techniques, including convolutional neural networks, transformer networks for spatial-temporal analysis, and unsupervised domain adaptation for both image-based and video-based tasks. Notably, in Chapter 3 and Chapter 4, there will be dedicated reviews for specific topics, such as partial domain adaptation, while Chapter 2 remains a general introduction for some commonly adopted research works.

Chapter 3 introduces the Multi-modality Cluster-calibrated partial Adversarial Network (MCAN), which uses multi-modal information to achieve better domain adaptation performance. The involvement of multi-modal information greatly helps the adaptation performance, as information from a single domain can be biased. The integration of multi-modal information also enhanced the estimation accuracy

of the target label space, making the filtration of redundant classes in the source domain more accurate and efficient. Moreover, as the key contribution, a calibration method is proposed to make identifying target classes even more accurate. This is achieved by utilizing the empirical patterns discovered when analyzing the high-dimensional features produced by deep models. In summary, MCAN effectively integrated multi-modal information for partial video domain adaptation and leveraged a novel calibration method to make the adaptation process more efficient and accurate.

Chapter 4 pointed out that while continuous domain adaptation is a commonly encountered scenario in video domain adaptation, it is largely unexplored. Therefore, in this chapter, this thesis first specifies a novel video domain adaptation scenario, namely Continuous Video Domain Adaptation (CVDA). In CVDA, it is assumed that a model needs to be continuously adapted to new arriving domains without being able to access historical data and supervision to emulate real-world scenarios closely. To enable better CVDA, this research proposal is to learn from samples attentively based on prediction confidence. This thesis proposes a novel data generalization enhanced self-knowledge distillation method to preserve learned knowledge. With both methods, this thesis proposes a Confidence-Attentive network with generalization enhanced self-knowledge disTillation (CART) as our main contribution. Extensive experiments show that CART can effectively preserve learned knowledge while learning based on noisy unsupervised information.

Chapter 5 summarizes the thesis and envisions some future works.

Chapter 2

Literature Review

Modern domain adaptation generally involves adapting a deep neural network with a set of adaptation techniques to a target domain. Therefore, this chapter reviews the related works of this thesis from two perspectives, i.e., backbone networks and adaptation techniques. Specifically, for backbone networks, a brief review of handcrafted feature-based methods is given, followed by reviews of recent progress in deep learning models for video action recognition, including convolutional neural networks and transformer networks. For adaptation techniques, this chapter reviews conventional Unsupervised Domain Adaptation (UDA) methods as well as recent advancements in Video-based Unsupervised Domain Adaptation (VUDA).

2.1 Video Action Recognition

Video action recognition involves analyzing a video with many frames rather than a single image, as is the case in image-based recognition tasks. This necessitates processing multiple frames from a video to generate features that can be utilized for classifying human actions. In the early stages of research, handcrafted features were predominantly used, and traditional machine learning algorithms such as Support Vector Machines (SVM) were employed for classification. However, with the rapid advancement of deep learning methods in recent years, handcrafted features have been largely supplanted by features generated by deep neural networks, which offer greater robustness to variance and improved scalability when given sufficient data. Presently, convolutional neural networks (CNNs) serve as the mainstream

approach for processing video data, though emerging studies suggest that self-attention-based transformer networks could become the next-generation standard for video feature extraction. This section provides a brief introduction to vision-based action recognition methods that utilize handcrafted features and offers a comprehensive review of vision-based action recognition techniques based on deep learning.

2.1.1 Handcrafted Visual Features

As a common practice in both machine learning and deep learning, feature extraction from raw data is essential for efficient classification since useful information is often implicitly represented within the raw data. Long before the advent of deep learning methods, researchers began utilizing various types of handcrafted features to analyze video clips and classify human actions depicted in them. Although these handcrafted features have been largely superseded by more robust features generated by neural networks, these early works provided valuable insights into the development of video understanding research. Handcrafted features can be broadly classified into several categories, including local/global feature descriptors, trajectory-based methods, and feature encoding and aggregation.

Local/Global Feature Descriptor Local feature descriptors primarily capture appearance and motion information in small regions of video frames, whereas global feature descriptors represent an entire video by considering a broader range of frames. Notable examples of local descriptors include Space-Time Interest Points (STIPs) [12–14] and Histogram of Oriented Gradients (HOG) [15]. For instance, STIPs-based methods employ the well-established Harris corner detector for image feature extraction to detect interest points in videos. Researchers extended the concept of the original Harris corner detector to detect rapid changes across both spatial and temporal dimensions. Following similar inspiration, the 3D-Hessian detector was proposed to find interest points in the spatiotemporal domain [16]. While various local feature descriptor methods have been proposed, they are often susceptible to changes in camera view angles and background movements, limiting their applicability in real-world scenarios.

Conversely, global feature descriptor-based methods like Motion History Image (MHI)[17] and Motion Energy Image (MEI)[17] focus on the entire video. MHI,

for example, computes and summarizes a single "motion image" to represent the motions occurring in a video clip. Likewise, MEI also generates a single image output, albeit without encoding temporal order. Although global descriptor methods incorporate more information from an entire video, they are limited by their inability to attend to specific changes between frames. These methods are also generally sensitive to occlusions, as they require clear human contours to function effectively.

Trajectory-based Methods Trajectory-based methods involve tracking points over time to capture motions as a set of trajectories. Two representative methods of trajectory-based methods are dense trajectories and improved dense trajectories[18]. These methods densely sample points in video frames and track them over time using optical flow. In iDT, for example, features such as HOG and the Motion Boundary Histogram (MBH) [19] were integrated. Later, works like SDT [20] were proposed to improve trajectory-based methods' robustness during camera movements. Overall, compared to other methods, trajectory-based methods are robust against camera view angle changes, but they require accurate human skeleton inputs, which can be challenging to acquire.

Feature Encoding and Aggregation This line of work largely focuses on combining local features into a compact and discriminative representation suitable for classification. Though conceptually similar to modern deep neural networks, the data flow is largely defined manually instead of being learned automatically by machines. Popular methods in this category include Bag of Visual Words (BoVW) [21], and Fisher Vector Encoding [22]. A general approach taken by these methods involves clustering local features into a visual vocabulary and then quantizing the features into a compact representation. Finally, the resulting compact representation can be used as input for SVM, and some advances were achieved. In summary, although comparatively better than simply classifying based on local features, this line of work is still largely limited by the fact that most of its components are handcrafted and cannot be optimized for vastly diverse real-world cases, making it unsuitable for industrial-grade applications.

Overall, handcrafted features have been used for many years and achieved promising results on smaller early datasets, including HMDB51 [23] and UCF101 [24]. However, they are largely incapable of handling diverse real-world cases due to the construction of features not always capturing important information from video

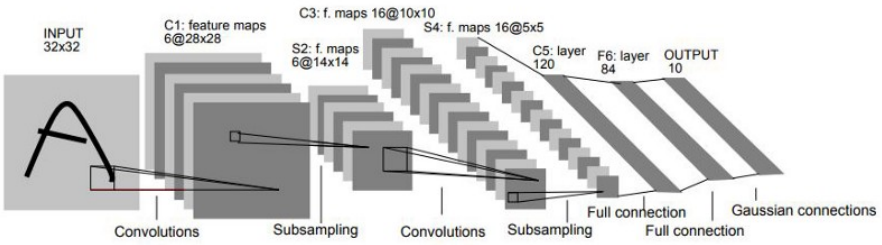

FIGURE 2.1: The structure of LeNet [1] where the input image is processed by a few layers of convolutional layers to extract features followed by fully connected layers to classify the extracted features.

clips, rendering them less applicable to larger and more challenging datasets such as Kinetics-400 [25].

2.1.2 Convolutional Neural Network

Starting around 2015, the rapid rise of deep learning-based methods began to outpace conventional handcrafted feature-based methods. Deep learning methods generally consist of a set of differentiable operations that can be optimized by empirical risk minimization (ERM) [26]. Among these, convolutional neural networks (CNNs) have been particularly successful in the areas of image classification and video action recognition.

Convolutional neural networks (CNNs) were first proposed by LeCun et al. [1], and their main idea is to learn convolutional kernels that are used to perform convolution on 2-dimensional inputs. By stacking such convolutional operations together, along with other necessary operations, including activation and fully connected layers, pioneering works like LeNet [27] have achieved great success in handwritten digit classification. Thanks to the rapid rise of GPU computing power, deeper and more complex CNNs emerged after the success of LeNet. These improved versions, such as AlexNet [28] and ResNet [29], achieved state-of-the-art performance on the ImageNet dataset [30] in 2012 and 2015, respectively.

Similar to the case of handcrafted features, convolutional neural networks for videos are largely adapted and improved variants of image-based convolutional neural networks. Some early works directly employed 2D CNNs as the backbone for feature extraction and temporal information is represented as a simple fusion of spatial features extracted from independent frames. Meanwhile, other works involved

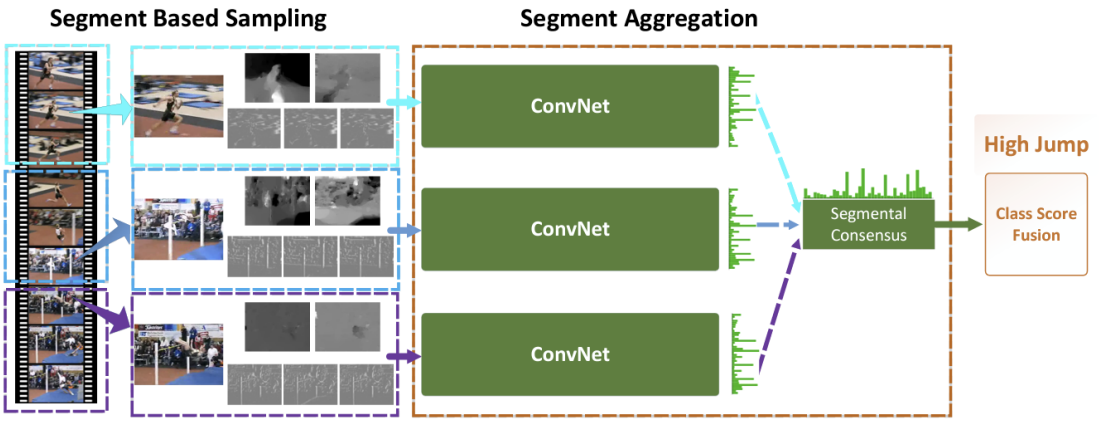

FIGURE 2.2: The structure of Temporal Segment Network (TSN) [2]. Video clips are divided into a few segments and frames are randomly sampled from those segments. Features from sampled frames are combined during the segmental consensus stage and the prediction is based on the output of the segmental consensus module.

leveraging a two-stream structure that processes RGB frames and optical flows in parallel. Later, 3-dimensional CNNs, also known as 3D CNNs, were proposed to efficiently integrate temporal information into compact spatiotemporal features. The following paragraphs of this section will dive deeper into the basics of both 2D CNNs and 3D CNNs for video action recognition.

2D-CNN based Methods. With the rise of deep learning [28], researchers began adapting CNNs to tackle video understanding challenges. One early work, DeepVideo [31], proposed using a single 2D CNN network to independently extract features from video frames and explored several relatively simple methods, such as early fusion, late fusion, and slow fusion, for combining these spatial features into spatiotemporal representations. However, it was found that transfer learning on UCF101 [24] was considerably less effective than handcrafted counterparts. Researchers also discovered that the network performed similarly regardless of whether the input was a single frame or multiple frames. These observations might indicate that the trained CNNs were not effectively capturing motion information, which could explain why deep learning-based methods were less effective for video problems compared to handcrafted feature-based methods.

Recognizing the shortcomings of CNNs, such as their inability to capture motion, optic flow [32] was later introduced to explicitly introduce motion information into CNNs' data flow. Optic flow is the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and

the scene [33]. In practice, optic flow is often represented as a two-channel image for each frame, where the first channel u corresponds to horizontal movement, and the second channel v corresponds to vertical movement. By removing redundant background information, optic flow simplifies the learning problem compared to using RGB images as inputs. This led to the development of two-stream networks [34], which achieved results comparable to previous state-of-the-art handcrafted feature-based methods for the first time, demonstrating that deep learning-based methods could surpass conventional handcrafted feature-based methods. This success spurred a wealth of follow-up research, further advancing video action recognition.

One seminal follow-up, Temporal Segment Network (TSN) [2], was proposed in the wake of the success of Simonyan et al. TSN experimented with various deep CNNs, including VGG [35], ResNet [29], and Inception [36], showing that deeper networks can achieve satisfying results on conventional datasets with appropriate training strategies, such as batch normalization and dropout. Specifically, TSN uses backbone networks such as ResNet [29] to extract frame-level spatial features, which are then fused using a consensus function. Although this structure is relatively simple, it achieves state-of-the-art performance on datasets and has inspired numerous follow-up works. For example, Temporal Relation Network (TRN) [4] was proposed to further enhance the effectiveness of the consensus function by more accurately representing temporal relations. Later works, such as TSM, are also built upon the structure of TSN, and they yielded promising improvements.

3D-CNN based Methods. Similar to the era of handcrafted features, researchers also attempted to extend 2-dimensional convolution to 3-dimensional convolution for processing video inputs. The introduction of 3D CNN first occurred in [37], although it was not deep enough to demonstrate its full potential. Later, Tran et al. [6] expanded upon this work and constructed a much deeper network called C3D, based on the basic structure of the VGG network. However, its performance was largely unsatisfactory, and significant improvements were needed. For instance, 3D CNNs are difficult to optimize, necessitating large-scale video datasets with diverse video contents and action categories. Training them also proved to be more challenging, as 3D CNN networks require significantly more memories and computations than their 2D counterparts due to their 3-dimensional kernels being naturally larger than 2-dimensional kernels. Another issue is that 2D CNNs can

fully utilize pre-trained weights, e.g., a pre-trained network on ImageNet [38], while 3D CNNs need to be trained from scratch.

This situation changed in 2017 with the introduction of I3D[7]. The main contribution of I3D was the revelation that by simply loading dilated pre-trained weights of a 2D CNN on image datasets, 3D CNNs could avoid repetitive training from scratch. The results achieved by I3D demonstrated that 3D CNNs could be as accessible as 2D CNNs while being more capable in terms of capturing temporal relations. Another work aiming to reduce the computational cost of C3D was R(2+1)D[39], which splits the 3-dimensional spatiotemporal kernel into one 2-dimensional spatial kernel and one 1-dimensional temporal kernel. This simplification maintained the performance advantages of 3D CNNs while significantly reducing the required computational resources. This strategy was also employed by S3D[40] and P3D[41]. Additionally, Channel-Separated Convolutional Network (CSN) [42] proposed performing convolution operations across channels separately, further alleviating the resource requirements of 3D CNNs. More recently, the SlowFast[43] network was proposed, which fuses information from a slow and fast pathway. The fast pathway is designed to process more frames but with limited channels, while the slow pathway processes fewer frames but with more channels. This ensures the simultaneous capture of long-range relations and important static features, achieving state-of-the-art results in many benchmarks.

In summary, 3D CNNs are powerful end-to-end methods for video action recognition, capable of working with only RGB input as they have the ability to accurately extract temporal relations. However, the structure of CNNs might not be the optimal choice for processing consecutive frames, as the advantage of CNNs lies more in their ability to understand 2D geometries rather than temporal relations. This is also partially supported by results obtained in this thesis, where CNNs can be confused by changes in the temporal characteristics of videos and be outperformed by two-stream methods in domain adaptation tasks.

2.1.3 Transformer Network

Transformer-based networks, a recent development in the field, are built upon a novel operation known as the self-attention mechanism. Originally designed for Natural Language Processing tasks, these networks quickly demonstrated their

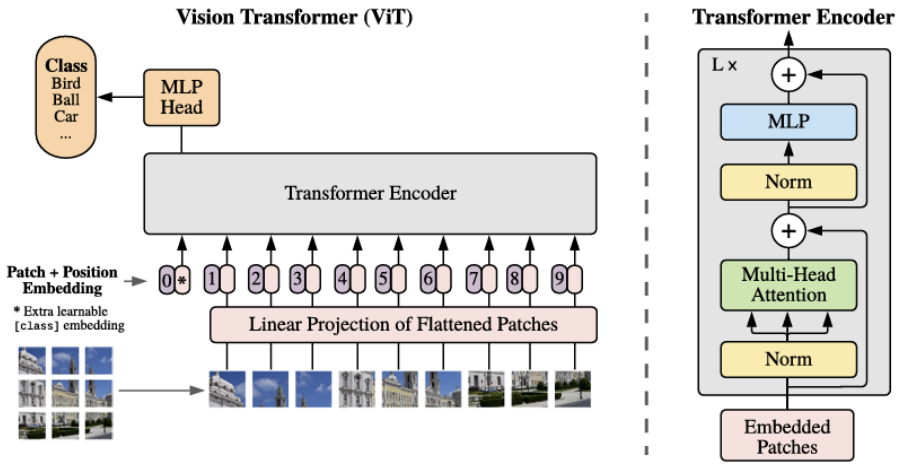

FIGURE 2.3: The structure of Vision Transformer (ViT) [3]. On the left side, it shows that a single image is represented by a grid of patches that are regarded as tokens for the transformer network to process. On the right side, it shows the structure of a single transformer block inside ViT.

potential for processing image and video data. The first work to adapt the self-attention mechanism for image classification was ViT [3], which suggested that transformer-based networks could be more scalable than conventional CNNs given sufficient data. Subsequently, video-based methods such as ViViT [44] and TimeSFormer [45] rapidly emerged. In transformer-based networks, multiple self-attention layers are stacked and connected to generate a representation for the input. Within each self-attention layer, the input is projected into three vectors: Query (Q), Key (K), and Value (V) using learnable transformation matrices. Multiplication is performed between Q and K to create attention maps, which are then further processed by multiplying with V to incorporate input information. Although the precise mechanism of self-attention remains a topic of debate, researchers generally agree that this structure allows networks to attend to different objects of interest in the inputs. For example, a network may learn to attend to human bodies in a video frame when tasked with classifying human actions.

In summary, transformer-based networks have started to exhibit significant potential in terms of generalizability and robustness. Research such as TimeSFormer [45] has also indicated that transformer networks are better suited to process temporal relations, which may be attributed to the inherent suitability of the self-attention mechanism for modeling temporal relations. However, as related research has only recently begun, it will take some time before the full potential of the self-attention mechanism can be exploited. To facilitate the understanding of transformer-based

networks, Chapter 4 of this thesis primarily focuses on transformer-based feature extraction networks, as they have proven to be both generalizable and consistent.

2.2 Domain Adaptation

Neural networks are primarily trained using empirical risk minimization, which implies that for networks to perform as expected on test data, the training data should share the same distribution as the test data. Otherwise, the test results can be significantly degraded due to sub-optimal predictions generated by the network. This discrepancy in distribution, commonly known as domain shift, indicates that the source domain (training dataset) has a different distribution than the target domain (test dataset). For instance, images captured by surveillance cameras and those taken by phones may differ in characteristics such as resolution, focus, and color, causing a model trained on surveillance camera data to be confused by inputs from phone cameras. To address this issue, domain adaptation is necessary. When sufficient supervised data from the target domain is available, a common approach for adaptation involves fine-tuning the model on the target dataset using standard training methods, such as empirical minimization. This straightforward strategy is widely adopted in various applications due to its promising results. However, in cases where collecting sufficient supervised data is challenging, fine-tuning may not be feasible. Consequently, Unsupervised Domain Adaptation (UDA) methods have been proposed in recent years to facilitate adaptation from the source domain to target domains without supervised data, reducing the cost of model adaptation. In this section, conventional image-based unsupervised domain adaptation methods are first introduced, followed by recent advancements in unsupervised domain adaptation methods for videos. This section also provides insights into the limitations of current video-based unsupervised domain adaptation methods in certain applications, which serves as motivation for this research to investigate specific domain adaptation scenarios for video recognition models.

2.2.1 Unsupervised Domain Adaptation

Since supervised domain adaptation requires supervised datasets which can be hard to obtain in real-world application scenarios, Unsupervised Domain Adaptation

(UDA) methods were proposed to avoid using supervised datasets. In UDA, the model trained on the source data is transferred to the target domain with the unsupervised target domain data and the essential goal of UDA methods is to minimize the domain discrepancy between the source and target domains. Previous works [46] have drawn the upper bound for the expected error $\epsilon_T(h)$ of the target samples and it is formulated as follows:

Theorem 2.1. *Let $\mathcal{H}$ be a hypothesis space and $\mathcal{U}_S, \mathcal{U}_T$ be samples drawn from distributions p_S of the source domain and p_T of the target domain respectively. For hypothesis $h \in \mathcal{H}$, the expected error of the target samples is bounded by:*

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + \mathbf{C}_0 + \lambda, \tag{2.1}$$

where $\epsilon_S(h)$ denotes the source error while $\frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T)$ is the empirical $\mathcal{H}\Delta\mathcal{H}$ divergence on samples $\mathcal{U}_S, \mathcal{U}_T$ drawn from distributions p_S, p_T. λ is the error of an ideal hypothesis for both source and target domains, while $\mathbf{C}_0$ is a constant term determined by the complexity of the hypothesis space $\mathcal{H}$. Notably, the second term is often approximated with discrepancy metrics in practice as the actual discrepancy is hard to be calculated directly.

According to Theorem 2.1, it shows that the key to decreasing the error rate on the target data is to minimize the second discrepancy term. This is because the first term, i.e., the source data error, is determined once trained on the source data, and $\mathbf{C}_0$ and λ are also determined by the selected model and λ is often considered small and negligible as an ideal hypothesis can perform equally well on all domains. Therefore, mainstream methods largely targeted minimizing the second discrepancy term, and two main approaches were proposed: adversarial-based methods and statistic-based methods.

Adversarial Domain Adaptation. Largely inspired by the success of Generative Adversarial Networks (GAN) [47], researchers proposed adversarial domain adaptation methods [8] that employ a feature extractor and a domain discriminator to minimize the domain discrepancy between the source and target domains. Specifically, such networks are trained in an adversarial manner: the domain discriminators learn to discriminate source features from target features, whereas the feature generators learn to confuse the domain discriminators. Eventually, the feature generator can learn to generate domain-invariant features such that the discriminator

is confused. As a result, since the domain discrepancy is effectively reduced during such a process, the discrepancy term in Theorem 2.1 is also decreased, resulting in better adaptation performance on the target domain. In practice, the domain discriminators are implemented as a stack of fully connected layers to distinguish source features from target features. The output of the feature generator is first fed to a Gradient Reverse Layer and then passed to the domain discriminator to achieve the goal of adversarial learning. Such structure soon becomes the mainstream method as it is plug-and-play and can achieve satisfying results on many datasets. Its success also inspired many other researchers and an ample of follow-ups were proposed. Consequently, PixelDA [48] was proposed to perform both feature-level and pixel-level domain adaptation. Additionally, CyCADA [49] introduced cycle loss and semantic consistency loss over both source and target classes on top of pixel domain adaptation. Later, works such as MDD [50], ADDA [51], and CADA [52] were proposed to further improve the performance of adversarial domain adaptation-based methods. In MDD, a discriminator gate [53] is introduced to filter negative transfer. In ADDA, the researcher performs discriminative representation learning first followed by the mapping of the target data to learn the representation space by asymmetric mapping. For CADA, the researchers further improve ADDA by fine-tuning the trained source feature generator while training the target feature generator via adversarial learning.

Statistics-based Domain Adaptation. As another viable domain adaptation approach, statistics-based domain adaptation methods seek to approximate the domain discrepancy based on statistics of the extracted features of the source domain and target domain data. In this manner, given a differentiable discrepancy loss, one can minimize this loss to achieve the goal of minimizing the discrepancy term in Theorem 2.1. The seminal work of such an idea is Maximum Mean Discrepancy (MMD) [9]. MMD operates based on the maximum mean discrepancy, which is a kernel-based statistical test used to measure the similarity between two probability distributions. During operation, the mean embeddings of the source and target domain samples in a high-dimensional Reproducing Kernel Hilbert Space (RKHS) are compared. The founding idea of MMD is that if two distributions are close in the RKHS, then the distributions can be considered similar. Therefore, given a deep neural network, one can first generate video features based on inputs and then calculate the mean maximum discrepancy as a loss term. By minimizing this loss, one will effectively achieve the goal of minimizing the domain discrepancy between

the source and target features and achieve efficient domain adaptation. However, the MMD is sensitive to the choice of kernel functions used to map data into RKHS and it requires a pairwise comparison between source and target samples, which may lead to higher computation costs given a larger unsupervised target dataset. Thus, a line of works based on MMD [9] was proposed to improve its effectiveness. MK-MMD [54], explored the idea of using multiple kernel functions, and Deep Adaptation Network [55] investigated enabling MK-MMD for deeper networks. Alternatively, other research suggests that one can also approximate the domain discrepancy between the source and target domain based on other types of metrics. For instance, SISS and SIE [56] are proposed to measure the discrepancy between source and target domains on the Riemannian manifold instead of RKHS, while CORAL is proposed to measure the discrepancy via covariance matrices.

In summary, the core objective of domain adaptation methods is well-defined by Theorem 2.1, i.e., minimize the domain discrepancy term. Adversarial-based methods took the way of leveraging adversarial networks to approach the goal of minimizing the discrepancy, while statistics-based methods employ a more direct way, i.e., approximate the discrepancy using certain discrepancy metrics, and minimize the metric to achieve minimization of domain discrepancy. Although UDA methods had achieved promising results over the years, researchers are beginning to find that sometimes the assumption made in the standard UDA problems can be restrictive. For example, UDA requires equal label space between the source and target domains. This can be hard to achieve, as identifying what classes are involved in the target domain can be exhaustive and costly. On the other hand, UDA methods also do not consider the adaptation process continuous, making it less applicable if the target data are arriving as batches and the model needs to continuously adapt to each arriving batch individually. Those shortcomings largely inspired the works being done in this thesis and our methods addressing those issues will be discussed in detail in Chapter 3, and Chapter 4.

2.2.2 Video-based Unsupervised Domain Adaptation

UDA researches [8, 9, 51, 52, 54] largely focus on image-based tasks as it is a typical case of domain adaptation. They can be applied to a variety of tasks, such

as image recognition, objection detection, as well as semantic segmentation. However, compared to image-based UDA research, Video-based Unsupervised Domain Adaptation (VUDA) works are of very limited quantity. This is partially due to the extraction of spatial-temporal features being much more challenging, and this was achieved in more recent years. In the early days, Waqas et al. [57] improved the generalization ability by decreasing the influences from the background, and Xu et al. [58] mapped source and target domains to a common feature space via shallow neural networks. More recently, Arshad et al. [59] take the approach of domain adversarial networks [8] to tackle VUDA problems, and Zhang et al. [60] approach VUDA with discrepancy-based methods. Later, TA^{3}N [10] proposes to leverage both domain adversarial networks and information-entropy-based attention mechanisms to tackle VUDA on larger datasets, e.g. [23, 24, 61] while ACAN [62] further applies the domain adversarial network to correlation information within videos. Naturally, to handle diverse videos, multi-modality VUDA networks are also proposed, and some of them integrate optic flow as the additional modality [63–66]. Qi et al. [66] propose a unified framework for multi-modal domain adaptation based on covariant multi-modal attention and multi-modal fusion module. MM-SADA [63] leverages RGB and optic flow to better recognize fine-grained human actions. In DLMM [65], a differentiated adversarial learning process is applied to different modalities, and teacher/student sub-models are applied to estimate the reliability of recognition results. Later, based on RGB and optic flow, Song et al. [64] propose to integrate contrastive learning to tackle DA problems. TCoN [11] instead proposes to align the distributions of video features using a novel cross-domain co-attention mechanism across the temporal dimension. Meanwhile, SAVA [67] proposed to add auxiliary tasks as part of the training goal. Overall, VUDA research focuses more on aligning spatial-temporal features, especially the temporal features, as it is the signature difference between video and image data. Despite its success, VUDA also faces some practical challenges, such as the need to align the label spaces and the challenge of continuous adaptation, as mentioned in Section 2.2.1. Those challenges provide important motivations for research, demonstrated in the following chapters.

2.3 Cross-Domain Video Action Recognition Datasets

With the rise of deep learning methods, the need for large and diverse datasets is becoming unprecedented as the performance of deep learning method often scale well as the dataset becomes larger and more diverse. Following such a trend, recent video action recognition datasets are becoming much larger than datasets used to train and test handcrafted features. To build a video action recognition dataset, a list of actions will be defined first. Then, videos are created either manually by researchers or collected from large-scale public platforms such as YouTube. Finally, the collected data will be cleaned and annotated. Notably, the annotation process often includes identifying relevant frames of action since raw video can be long and much irrelevant information can introduce extra noise. In the following paragraphs, some mainstream datasets that will be used in this thesis to crate challenging domain adaptation datasets are introduced.

HMDB51 [23] was introduced in 2011 as a pioneering dataset for video action recognition research. Its video clips are largely collected from movies, and some are from public databases such as Prelinger archive, YouTube, and Google videos. In total, HMDB51 contains 6,849 clips and 51 action categories. Officially, HMDB51 has three splits for training and testing and most previous works report the top-1 accuracy either on split 1 or average across all three splits.

UCF101 [24] was introduced in 2012 as an extension for the previous UCF50 dataset. It contains 13,320 videos from YouTube and a total of 101 classes are contained. Similarly, UCF101 also has three official splits and the accuracy can be reported in a similar manner to HMDB51.

Sports1M [68] was introduced in 2014 as the first large-scale video action dataset, which contains more than 1 million YouTube videos annotated with 487 sports classes. The action categories are fine-grained, making it challenging to correctly predict the action class between similar classes. It has an official 10-fold cross-validation split for evaluation.

Moments in Time [69] is a dataset first proposed in 2018 and it is a large-scale dataset designed for event understanding. Video clips are all 3 seconds clips, and there are 1 million video clips in total spread across 339 classes. What makes this

dataset slightly different from other datasets is that video clips in this dataset also involve people, animals, objects, and natural phenomena, making it vastly different from ordinary datasets that solely concentrate on human-based actions.

Kinetics-400/600/700 are a series of large-scale datasets that are widely adopted in many research works as the involved video clips are diverse and of large quantity. Kinetics-400 [25] was first publicized in 2017, and it consists of around 240k training and 20k validation video clips that are manually trimmed down to 10 seconds to exclude redundant contents. As the name suggested, there are a total of 400 classes in Kinetics-400. In 2018 and 2019, the authors further expanded the Kinetics-400 to Kinetics-600 and Kinetics-700. Today, Kinetics series datasets have become the standard dataset for testing the performance of a deep action recognition model as the data diversity and quality is above average.

ARID is a special dataset built for the purpose of testing video action recognition models in dark environments. Normal datasets largely consist of samples shot in normal lighting conditions and the model trained on them can be prone to the change of such conditions. It features 5,572 clips with more than 320 clips per action, with a total of 11 actions. Notably, ARID is created solely by volunteers in 24 scenes. Overall, the ARID dataset offers a unique chance to test the generalization ability of video action recognition models in extreme environments, making a decent choice to set up a cross-domain adaptation benchmark.

In the cross-domain adaptation context, a common practice to create a domain adaptation benchmark dataset is to identify a few commonly shared classes across different mainstream datasets, such as HMDB51 [23], UCF101 [24], and Kinetics-400 [25], and regard one dataset as the source domain while using another dataset as the target domain. Following this pipeline, benchmark datasets can be created according to the needs of each research work. Some common formulations are the UCF-Oplympic, UCF-HMDB$_{\text{small}}$ in TA^{3}N [10], Epic-Kitchen-DA from Epic-Kitchen project [70], etc. In summary, modern domain adaptation datasets for video action recognition are largely built using mainstream large-scale datasets for video action recognition. Since there are numerous shared classes, great flexibility is allowed for testing many different kinds of adaptation scenarios, such as standard Video-based Unsupervised Domain Adaptation, Partial Video Domain Adaptation, Continuous Video Domain Adaptation, as well as other types of domain adaptation scenarios.

2.4 Conclusion

This provides a comprehensive survey of vision-based action recognition methods, including handcrafted feature-based methods, convolutional neural networks-based methods, and self-attention transformer-based methods. The fundamental mechanisms behind each method are discussed and highlighted their advantages and disadvantages against action recognition benchmarks. This also provided us with key motivations that lead to the research that will be discussed in this thesis, i.e., current deep models for vision-based action recognition may perform poorly in practical domain adaptation scenarios due to significant domain shifts. Meanwhile, a brief introduction is provided for Unsupervised Domain Adaptation (UDA) methods. The founding theory of UDA is displayed, and some insights are also provided. Then, two mainstream approaches of UDA, i.e., adversarial domain adaptation and statistic-based domain adaptation, are also introduced. Lastly, prevailing large-scale datasets for vision-based action recognition are introduced, and those datasets will be employed to formulate new cross-domain benchmarks in the following chapters of this thesis.

Chapter 3

Multi-Modality Partial Video Domain Adaptation

3.1 Introduction

Though video action recognition has been studied for years, one key challenge of applying it in the real world is that domain shifts between datasets would reduce the model performance across different video domains. Video-based Unsupervised Domain Adaptation (VUDA) methods [10, 11, 57–60] are therefore proposed to mitigate the domain shifts. While many studies achieve notable improvements, they generally assume the source label space and target label space are identical. Such an assumption is sometimes impractical as manually aligning the label spaces of different datasets can be tedious or impossible. In view of this, following the definition of Partial Domain Adaptation (PDA) [71], Xu et al. [72] propose to assume that the source label space subsumes the target one and define such scenario as Partial Video Domain Adaptation (PVDA).

As the main difference from VUDA, the existence of source-only outlier classes can bias the network in PVDA in such a way that target features may be misaligned to outlier classes. In such cases, the overall performance may be compromised due to the negative transfer of irrelevant source data triggered by the learned knowledge of outlier classes. To address the interleaving challenge of mitigating domain shifts and suppressing the negative transfer of irrelevant source data, attempts have been made in existing works [72] to filter out outlier classes using a class weight

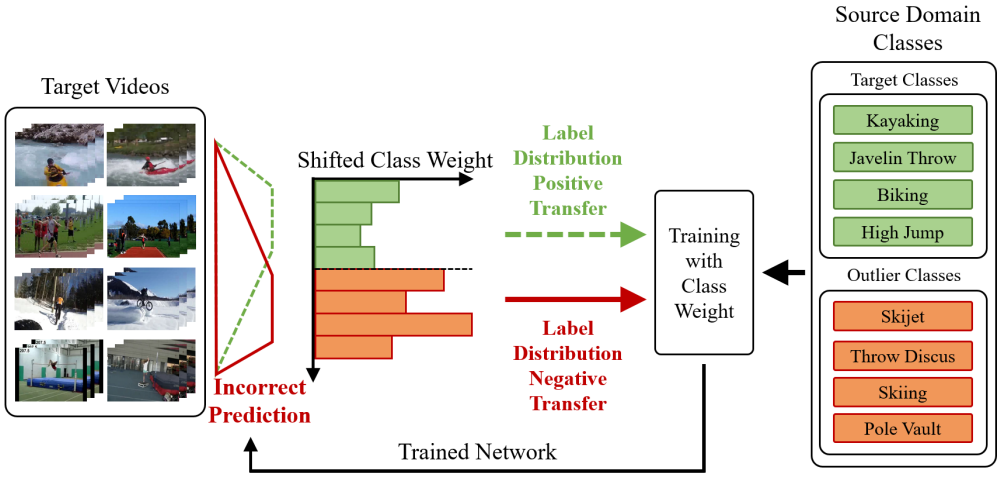

FIGURE 3.1: An illustration showing the existence of source-only outlier classes can cause negative transfers in PVDA. Bars in the class weight plot and classes in the source domain are placed following the same order. Target classes and outlier classes are chosen to be similar to each other and placed by order. Class weights are obtained by aggregating network predictions of target videos. Negative transfer of irrelevant source data is triggered when the network incorrectly aligns target videos to source-only outlier classes. Consequently, label distribution negative transfer arises and biases the network towards outlier classes. Best viewed in color.

that is enhanced by RGB-only multi-scale temporal pooling [4] and an attention mechanism. However, the utilization of RGB-only features is sub-optimal as the fused spatial features implicitly imply the presence of motions, rendering them insufficient to represent complex motions. Therefore, it is more desirable to integrate additional modalities, such as optic flow, to capture motions more effectively. With this objective in mind, this chapter proposes the construction of the Multi-modality partial Adversarial Network (MAN) by combining RGB and optic flow features to improve motion representation. Both types of features are constructed using multi-scale temporal pooling [4] to further enhance the representation of motion. In summary, the temporal feature extraction is efficiently enhanced, and a better class weight is composed to facilitate the suppression of negative transfer.

While it is the consensus in many studies [71–76] that the negative transfer of irrelevant source data in PDA and PVDA can be addressed by identifying outlier classes and having them filtered out, this research asks an obvious but less attended question: *what if the obtained outlier identification is incorrect?* In the case of class weight, it is termed as *label distribution negative transfer* since the learned label distribution that shifts away from the real target label distribution could

severely deteriorate the network performance. More specifically, such negative transfer is the cycle of learned incorrect predictions compromising the class weight and the class weight further confusing the network predictions. To circumvent this, this thesis argues that not all predictions should contribute equally to the class weight. In other words, the class weight should be calibrated such that incorrect predictions are down-weighted and correct predictions are up-weighted. Therefore, the question converts to the effective retrieval of prediction correctness of the target domain without target labels in PVDA. Inspired by previous works [77–79], this research pointed out that clustering can be leveraged to approximate such target supervision. Therefore, this chapter proposes to approximate the prediction correctness by exploiting cluster structures of video features. Lastly, by combining the class weight calibration method with MAN, this chapter constructs *Multi-modality Cluster-calibrated partial Adversarial Network (MCAN)*. In MCAN, the MAN and class weight calibration further complement shortcomings of each other, and PVDA performance uplift is achieved for tasks with more significant domain shifts.

In summary, our contributions are listed as follows:

- To improve the effectiveness of extracted temporal features, this chapter proposes MAN that utilizes multi-modal features and multi-scale temporal pooling to enhance class weight and network predicting performances.
- To mitigate the label distribution negative transfer, approximate the correctness of predictions is approximated using cluster structures of video features and weighing them accordingly. As a result, the shifted class weight is further calibrated such that it promotes positive transfers of relevant source data and suppresses negative transfers of irrelevant source data simultaneously.
- With the joint efforts of MAN and class weight calibration, MCAN achieves state-of-the-art performance on current prevailing benchmarks, and it is also the first study in the field of PVDA to exploit cluster structures. Thorough ablation studies and empirical analysis of MCAN additionally show that the class weight calibration is not sensitive to parameter initialization and can be applied to different networks.

3.2 Related Work

Video-based Unsupervised Domain Adaptation. Recently, many domain adaptation (DA) methods have been proposed to improve the network generalization ability on target domains. Most studies focus on image-based DA tasks [8, 50, 52, 54, 80, 81, 81–84], and only a few studies [10, 11] the Video-based Unsupervised Domain Adaptation (VUDA) problem. Viewers may refer to the previous Section 2.2.1 for a comprehensive understanding of those domain adaptation methods.

Partial Video Domain Adaptation. To closely emulate real-world scenarios, researchers have extended domain adaptation with many other definitions such as Partial Domain Adaptation (PDA), Multi-source Domain Adaptation [85], and so on. Specifically, PDA focuses more on how to identify target classes and source-only outlier classes. Such a challenge exists because that PDA allows the source label space to subsume the target one, and the network can misalign target features to outlier classes. Many image-based works [71, 73–76] approach this challenge from different aspects: SAN [74] selects out these outlier classes with a multi-discriminator domain adversarial network, and a weighting mechanism, IWAN [73] chooses to derive the probability of a source example belonging to the target domain, PADA [71] weighs each class by a class weight vector obtained as the aggregation of target predictions, ETN [75] identifies outlier classes by quantifying transferability of examples, and A^{2}KT [76] takes a progressive approach to gradually filter out outlier samples. Though PDA is well studied on image-based tasks, Partial Video Domain Adaptation was proposed more recently by Xu et al. [72], and many of these image-based PDA methods are incapable of handling PVDA tasks because video features are more complex and less separable. Based on PADA [71], PATAN [72] leverages multi-scale temporal pooling [4] and information-entropy-based attention mechanism to compute better class weights and filter out outlier classes more thoroughly. Nevertheless, PATAN [72] can still fall short of fitting the target label distribution when significant domain shifts exist. To further improve the accuracy and robustness of the class weight, the integration of multi-modal features into multi-scale temporal pooling is proposed, thereby enhancing the performance of PVDA with better class weights.

Clustering-based Domain Adaptation. In the field of unsupervised learning, clustering is a powerful tool to characterize high-dimensional features and retrieve supervised information. Thus, cluster structures are exploited in much Unsupervised Domain Adaptation (UDA) studies to mitigate domain shifts. The key idea is to acquire extra target supervision by exploiting cluster structures. For instance, MSTN [79] proposes to align target class centroids with source class centroids so that target features are more semantically aligned with the source features. Later, CAT [86] improves upon [79] by installing a teacher network to produce pseudo labels. Despite performing alignments between classes, DIRT-T [87] leverages the cluster assumption [88] to refine the decision boundaries of the classifiers. And SHOT [77] exploits the cluster structures directly by filtering out outlier classes in PDA tasks depending on the total of members in each cluster. In sum, clustering offers rich distributional information about the target features, and existing studies have exploited it from many different aspects. However, in PVDA, few studies have leveraged cluster structures to facilitate the promotion of positive transfers and suppression of negative transfers simultaneously. To further enhance the PVDA performance, it is believed that clustering can be applied and yield benefits. Therefore, a novel class weight calibration method is proposed in this research, which approximates the correctness of predictions and incorporates them into the class weight. This calibration method aims to suppress the negative transfer caused by label distribution and, as a result, facilitates positive transfers of relevant source data while suppressing negative transfers of irrelevant source data simultaneously.

3.3 Methodology

Similar to Video-based Unsupervised Domain Adaptation (VUDA), Partial Video Domain Adaptation (PVDA) is provided with the source domain with n_s labeled samples and source label space $\mathcal{C}_s$ $\mathcal{D}_S = \{(V_i^s, y_i^s)\}_{i=1}^{n_s}$ and target domain $\mathcal{D}_T = \{(V_i^t)\}_{i=1}^{n_t}$ of n_t unlabeled samples with target label space $\mathcal{C}_t$. What makes PVDA different from VUDA is it assumes that $\mathcal{C}_t \subset \mathcal{C}_s$ instead of $\mathcal{C}_s = \mathcal{C}_t$. The change of label space assumption means the outlier label space $\mathcal{C}_o = \mathcal{C}_s \backslash \mathcal{C}_t$ would exist and cause irrelevant source data to be negatively transferred. Thus, in PVDA, besides mitigating domain shifts, promoting positive transfers of relevant source

data and suppressing negative transfers of irrelevant source data is crucial. To mitigate the domain shift, the current mainstream method is to apply domain adversarial networks [10, 11, 72]. The domain adversarial network [8] is analogous to the Generative Adversarial Network (GAN) [47] as it forms a min-max game between the feature extractor and domain discriminator. To suppress the negative transfer of irrelevant source data, the key method is to obtain a class weight γ every s steps to down-weigh outlier classes [71] and have them filtered out. γ is obtained as the aggregation of all target predictions and can be viewed as a rough approximation of the real target label distribution. To improve the approximation accuracy, recent advances [72] leverage RGB-only multi-scale temporal features and information-entropy-based attention mechanism to enhance the class weight. Nevertheless, it is observed that RGB-only features can be insufficient to represent temporal information, e.g., motion, and additional modalities should be integrated. Following this inspiration, Multi-modality partial Adversarial Network (MAN) is proposed to generate more robust and transferable features while suppressing the negative transfer of irrelevant source data.

On the other hand, while computing class weight [71] is one of the simplest yet most effective Partial Domain Adversarial (PDA) methods [71, 73–76] for PVDA, the class weight is far from a perfect approximation of the target label distribution. Typically, the class weight would shift away from the real target label distribution due to incorrect predictions, and this causes the accidental down-weighing of target classes and up-weighing of outlier classes. In this thesis, this is termed as *label distribution negative transfer* because inaccurate class weight can cause the cycle of learned incorrect predictions to compromise the class weight and the class weight further confusing the network predictions. Moreover, it is particularly noted that the label distribution negative transfer includes the case where target classes are accidentally down-weighted while previous works attend less to this issue [10, 72, 77]. To this end, inspired by previous works [77, 79, 86, 87], this chapter proposes to suppress the label distribution negative transfer via calibrating the shifted class weight. Specifically, this chapter first studies the cluster structures of video features in PVDA tasks. Then it is proposed to up-weigh correct predictions and down-weigh incorrect predictions where the correctness of predictions is approximated by exploiting cluster structures in a novel way. Lastly, with the simultaneous promotion of positive transfers and suppression of negative transfers, the combination of class weight calibration with MAN leads to the Multi-modality Cluster-calibrated

partial Adversarial Network (MCAN). The thesis begins by revisiting PVDA, followed by a detailed illustration of MAN and the algorithms employed for the class weight calibration of MCAN.

3.3.1 PVDA with Adversarial Network

The main challenge of PVDA is to mitigate domain shifts and the negative transfer of irrelevant source data. The key idea to tackle them simultaneously is to learn domain-invariant features and filter out outlier classes with class weight γ. To obtain domain-invariant features, domain adversarial network [8] proposes to form a min-max game similar to Generative Adversarial Networks [47]. To filter out outlier classes, γ is obtained by aggregating network predictions, and the domain adversarial network is also weighted by γ to filter out outlier classes more thoroughly. The overall objective for prior PVDA networks [72] is formulated as:

$$\begin{aligned}\mathcal{L} = &\frac{1}{n_s}\sum_{i=1}^{n_s}\gamma[L_y(G_y(G_f(V_i^s)), y_i) - \alpha L_d(G_d(G_f(V_i^s)), d_i)] \\ &- \frac{\alpha}{n_t}\sum_{i=1}^{n_t} L_d(G_d(G_f(V_i^t)), d_i),\end{aligned} \tag{3.1}$$

where G_y is the source classifier, G_d is the domain discriminator, G_f is the feature extractor, V_i^s, V_i^t is the sampled video input from the source and target domain, y_i is the ground truth class label, d_i is the ground truth domain label, α is a trade-off hyperparameter to balance label and domain classification, and L_d, L_y are implemented as cross-entropy losses. For the class weight γ, it is obtained every s mini-batches as: $\gamma' = \frac{1}{n_t}\sum_{i=1}^{n_t}\delta(\hat{y}_i)$, where $\hat{y}_i = G_y(G_f(V_i^t))$ is the network prediction, and δ is a softmax function. In practice, γ' is additionally normalized by dividing its mean $\overline{\gamma'}$, i.e. $\gamma = \frac{\gamma'}{\overline{\gamma'}}$.

3.3.2 Multi-Modality Partial Adversarial Network (MAN)

Existing works [72] primarily focus on leveraging RGB-only features to obtain better temporal features and thus compute more accurate class weight. This can be insufficient as motion features are implicitly embedded in the temporally pooled [4] RGB-only features. To explicitly represent motion features, MAN is proposed as

a means to achieve this objective. The corresponding details of MAN are described below.

This chapter denotes the ith video input that contains multi-modal frames as $V_i = \{(v_{1,1}, v_{1,2}, \ldots, v_{1,M}), \ldots, (v_{N,1}, v_{N,2}, \ldots, v_{N,M})\}$, where $v_{j,m}$ is the jth frame of the mth modality and N, M is the total number of sampled frames and modalities, respectively. For the following equations that contain multi-modal features, the same sub-scripting rule is also applied. Thus, the main objective of MAN is formulated similarly to Eq.3.2 as:

$$\begin{aligned}\mathcal{L} = \frac{1}{n_s}\sum_{i=1}^{n_s}\gamma[L_y(G_y(F(\sum_{m=1}^{M}G_{f,m}(V_{i,m}^s))), y_i) \\ - \alpha\sum_{m=1}^{M}L_d(G_{d,m}(G_{f,m}(V_{i,m}^s)), d_i)] \\ - \frac{\alpha}{n_t}\sum_{i=1}^{n_t}\sum_{m=1}^{M}L_d(G_{d,m}(G_{d,m}(V_{i,m}^t)), d_i),\end{aligned} \tag{3.2}$$

where F is a fusing function for multi-modal features. In practice, F is implemented as MLP layers to align dimensions and element-wise addition to fuse features. Consequently, γ is obtained every s mini-batches as:

$$\gamma' = \frac{1}{n_t}\sum_{i=1}^{n_t}\delta(G_y(F(\sum_{m=1}^{M}G_{f,m}(V_{i,m}^t)))), \tag{3.3}$$

where the raw class weight γ' will be additionally divided by its mean. In practice, while RGB is often the first modality to be applied, optic flow is considered the second modality in many studies [2, 6, 7, 43]. This is largely due to the optic flow can explicitly express the motions between given frames and suppress irrelevant background information. Conventionally, generating optic flow requires the TV-L1 [89] algorithm to run offline, and each frame in the original video will be assigned a post-adjacent optic flow frame. Such an extraction pipeline is inflexible and consumes storage space. Moreover, the TV-L1 algorithm also falls short of handling long-range dependencies and occlusions. Given this, this chapter proposes to directly obtain optic flow frames between sampled frames online, i.e. during training, using learned neural networks. In this thesis, segment-based sampling is applied,

and the optic flows can be extracted as:

$$V_{i,2} = \{\mathcal{O}(v_{1,1}, v_{2,1}), \mathcal{O}(v_{2,1}, v_{3,1}), \ldots, \mathcal{O}(v_{N-1,1}, v_{N,1})\}, \tag{3.4}$$

where $\mathcal{O}$ is the learned neural network, sampled RGB frame $v_{j,1}$ is randomly obtained from the jth segment of the video and temporally ordered in $V_{i,1}$. For the following equations in this chapter, they also define RGB as the first modality and optic flow as the second, i.e. $m = 1$ for RGB and $m = 2$ for optic flow.

Further, the Temporal Relational Module [4] is applied to both modalities to enhance the motion representation. Based on the segmental sampling, the multi-modal frame-level feature $\mathbf{f}'_i$ for V_i can be denoted similarly as:

$$\mathbf{f}'_i = \{(f_{1,1}, f_{1,2}), (f_{2,1}, f_{2,2}), \ldots, (f_{N,1}, f_{N,2})\}, \tag{3.5}$$

where $f_{j,m}$ is the extracted feature using $G_{f,m}$. Next, multiple frame features $f_{j,m}$ in $\mathbf{f}'_i$ are combined to form the clip with r temporally ordered and randomly sampled frames where $r{\in}[2, N]$. Formally, the multi-scale feature $\mathbf{f}_i$ for V_i can be denoted as:

$$\mathbf{f}_i = \sum_{m=1}^{M} \sum_{r=2}^{N} \sum_{l=1}^{L} g_r(f_{1,m}, f_{2,m} \ldots, f_{r,m})_l \tag{3.6}$$

where $f_{j,m}$ here instead refer to the jth within the clip, r features are sampled from $\mathbf{f}'_i$ for the lth clip, L is the maximum number of clips $L = max(C_N^r, 5)$, g_r is implemented as Multi-Layer Perceptron (MLP) to fuse temporally concatenated frame features $f_{j,m}$. Overall, all clips with the same length are grouped as a single temporal scale, i.e. an r-frame relation.

3.3.3 Multi-Modality Cluster-Calibrated Partial Adversarial Network (MCAN)

To further mitigate the label distribution negative transfer that leads to deteriorating class weight, not all predictions should contribute to the class weight equally. To achieve this, MCAN is proposed, which incorporates a lightweight and effective calibration module. This calibration module is designed to enhance the PVDA performance further, building upon the effective design of MAN. Existing works [72] weigh the predictions by measuring their certainty with information entropy

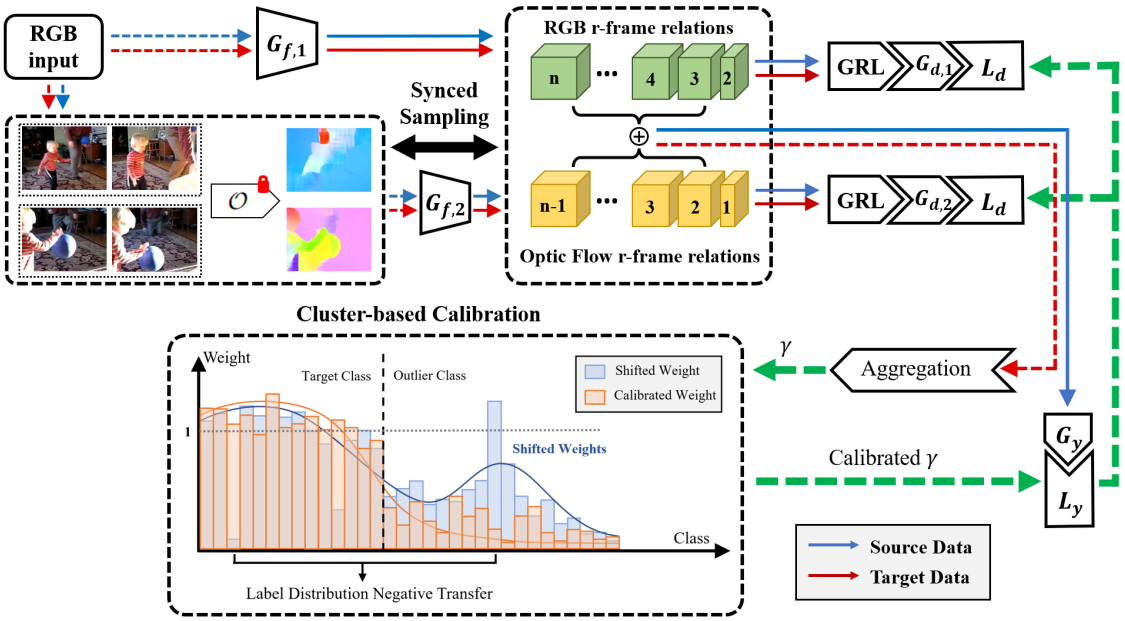

FIGURE 3.2: Illustration of the proposed MCAN architecture. $G_{f,1}, G_{f,2}$ is the feature extractor for RGB and optic flow modality; $\mathcal{O}$ is the optic flow estimation network; GRL is the Gradient Reversal Layer [4]; Optic flows are generated from sampled frames in each segment in the video. Calibrated γ is fed to weigh L_d, L_y. The dotted lines represent that the data flow is single-ended without backpropagation. The black solid arrow for Sampling Sync passes the segmental sampling information to $G_{f,1}, G_{f,2}$ and $\mathcal{O}$. Parameters of $\mathcal{O}$ are completely frozen and marked by the red 'lock'. Best viewed in color.

or directly discard identified outlier classes using cluster structures [77]. Though effective, their dedicated strategies may not be optimal for cases where the domain gap is large. Therefore, this section first proposes a lightweight and simple weighing strategy suitable for networks like MAN as follows:

$$\gamma' = \frac{1}{n_t} \sum_{i=1}^{n_t} \delta(\hat{y}_i) \omega(1 - H_{ent}(\delta(\hat{y}_i))), \tag{3.7}$$

where $H_{ent}(p) = -\sum_k p_k log(p_k)$ measures the certainty levels, and p_k is the probability of a sample being classified to the kth class. γ' should be divided by its mean $\overline{\gamma'}$ to obtain γ. However, leveraging information entropy to produce weights is not ideal. For instance, given two classes, an incorrect prediction $\hat{y}_1 = [0, 1]$ would produce just the same weight as a correct prediction $\hat{y}_2 = [1, 0]$. Furthermore, though correctly promoting positive transfers of relevant source data is important, entropy-based calibrating is particularly incapable of enabling such promotions. For example, given a class weight of three target classes $\gamma = [1.2, 0.8, 2.0]$, while

the predictions for the second class can be rather uncertain, entropy-based weighing is likely to produce a worse one, e.g. $\gamma = [1.5, 0.6, 3.1]$, due to the nature of information-entropy-based weighing is to down-weigh uncertain predictions. With the goal of promoting positive transfers and negative transfers simultaneously in mind and inspired by previous works [77, 79, 86, 87], this chapter proposes exploiting cluster structures of video features to suppress the label distribution negative transfer.

To exploit cluster structures, it is necessary to identify general patterns within video feature clusters. The focus of this thesis is on the utilization of the k-means algorithm, although alternative clustering approaches can be employed. More precisely, the problem of learning a $d \times k$ centroid matrix C and obtaining cluster assignments u_i for each video feature $\mathbf{f}_i$ is addressed by the k-means algorithm [90].

$$\min_{C \in \mathbb{R}^{d \times k}} \frac{1}{n_t} \sum_{i=1}^{n_t} \min_{u_i \in \{0,1\}^k} ||\mathbf{f}_i - C u_i||_2^2 \; s.t. \; u_i^\top 1_k = 1. \tag{3.8}$$

Solving the problem provides a set of optimal cluster assignments $(u_i^*)_{i<n_t}$ and the centroid matrix C^*. The centroids are obtained following [91] and the optimization of Eq.3.8 is performed following [92]. Upon observation, it is observed that with large domain shift and rather small K values:

- Features of incorrect predictions are more likely to be located closer to centroids.
- Features of incorrect predictions are likely to form large clusters, and this becomes more obvious when there are more source classes and samples.
- Correct predictions with separable features are more likely to form independent smaller clusters, and they can be overly up-weighted because they are certain and numerous.

To exploit these observations, a counter-intuitive but effective way would be down-weighing predictions whose features are close to centroids. Moreover, to reduce the excessive up-weighing of well-classified classes, a cluster-relative weighting scheme would be more sensible. In line with these conclusions: a novel weighting scheme

can be written as:

$$H_{cls}(\mathbf{f}_i) = \begin{cases} a & \tau \in (-\infty, \mu - \sigma) \\ \frac{(\tau-(\mu-\sigma))(b-a)}{2\sigma} + a & \tau \in [\mu - \sigma, \mu + \sigma) \\ b & \tau \in [\mu + \sigma, \infty) \end{cases} \tag{3.9}$$

where $a < b$, τ is the distance between $\mathbf{f}_i$ and its corresponding centroid Cu_i, i.e. $\tau = ||\mathbf{f}_i - Cu_i||_2^2$, μ is the mean of all τ in cluster $C^*u_i^*$, and, similarly, σ is the standard deviation.

Compared to entropy-based calibration [10, 72] and filtering out outlier classes in hard ways [77], our weighing strategy has these benefits:

- It roughly achieves the same positive effects as these methods.
- It partially avoids up-weighing incorrect predictions like entropy-based calibration.
- It is more robust and flexible than discarding outlier classes directly.
- It partially avoids overly up-weighing well-classified target classes since the weighting scheme is relative to cluster structures.

Last, to make cluster-based calibration more broadly applicable, the final weight is an addition of both cluster-based calibration as well as entropy-based calibration.

$$\gamma' = \frac{1}{n_t} \sum_{i=1}^{n_t} \delta(\hat{y}_i)(\beta\omega(1 - H_{ent}(\delta(\hat{y}_i))) + H_{cls}(\mathbf{f}_i)), \tag{3.10}$$

where β balances both calibrations, ω is set to align the range of H_{ent} to H_{cls}, and γ is obtained by dividing γ' with its mean $\overline{\gamma'}$.

By applying Eq.3.10 to MAN, MCAN is obtained. In sum, MCAN (1) leverages the explicitly expressed motion features in optic flow to form a multi-modality network and improves the overall PVDA performance, and (2) it particularly addresses the issue of label distribution negative transfer by facilitating the positive transfer of relevant source data and suppressing the negative transfer of irrelevant source data simultaneously. The illustration of the overall architecture is presented in Fig.3.2

3.4 Experiment

In this section, experiments on three benchmarks are conducted to evaluate MCAN against previous PVDA/VUDA methods and several image-based PDA methods. The comparisons between our methods and other methods are first shown, followed by thorough empirical analysis and ablation studies.

3.4.1 Setup

The evaluation of the network performance is conducted on the following PVDA benchmarks: UCF-HMDB$_{partial}$, MiniKinetics-UCF, and HMDB-ARID$_{partial}$ introduced by [72]. Thereinto, ARID(**A**) [93] dataset is particularly created for video shot in darkness and has larger domain shifts from common datasets, including HMDB51(**H**) [23], UCF101(**U**) [24], MiniKinetics(**M**) [94], etc.

UCF-HMDB$_{partial}$. The UCF-HMDB$_{partial}$ dataset is constructed upon 14 shared classes between UCF101 and HMDB51 datasets. A total of 2780 videos are sampled, and the first 7 classes in alphabetical order are chosen to be the target classes. The original split of training and evaluation is maintained. In sum, UCF-HMDB$_{partial}$ offers settings: **U-14→H-7** and **H-14→U-7**.

MiniKinetics-UCF. Similarly, the MiniKinetics-UCF dataset contains 45 shared classes and 18 target classes for the target domain. Across MiniKinetics-200 [94] and UCF101 [24], a total of 22,102 videos are contained, which is much more than other benchmarks. Such a scale also suggests that MiniKinetics-UCF can be more qualified to represent real-world scenarios. Overall, MiniKinetics-UCF offers settings: **M-45→U-18** and **U-45→M-18**.

HMDB-ARID$_{partial}$. The ARID [93] dataset was created based on videos shot in darkness. Therefore, transferring to ARID [93] from other datasets, e.g. HMDB51 [23], can be challenging. Overall, 10 source classes are selected, and 5 of them are chosen as target classes using the same protocol applied to UCF-HMDB$_{partial}$ and MiniKinetics-UCF101. This yields **H-10→A-5** and **A-10→H-5**.

For implementation, PyTorch [95] library is used to implement all domain adaptation methods. For RGB modality, TSN [2] trained on ImageNet [38] is utilized. RAFT+GMA [96] trained on Sintel [97] is leveraged as the optic flow estimation

TABLE 3.1: Results of MAN and MCAN on three pairs of datasets

Methods	UCF-HMDB$_{partial}$		MiniKinetics-UCF		HMDB-ARID$_{partial}$	
	U-14→H-7	**H-14→U-7**	**M-45→U-18**	**U-45→M-18**	**H-10→A-5**	**A-10→H-5**
TRN [4]	62.85%	78.95%	88.57%	64.30%	23.33%	26.00%
DANN [8]	60.95%	74.44%	85.94%	64.06%	24.10%	34.00%
TA^{3}N [10]	50.49%	70.68%	75.70%	48.23%	18.30%	24.00%
PADA [71]	65.71%	82.33%	89.45%	63.35%	24.36%	34.00%
ETN [75]	67.88%	82.89%	83.33%	63.59%	19.49%	28.82%
MK-MMD [54]	58.57%	82.71%	87.85%	63.82%	26.67%	34.67%
MCD [81]	55.71%	73.31%	88.58%	65.72%	19.74%	33.33%
MDD [50]	62.58%	80.45%	85.79%	66.66%	25.13%	26.00%
PATAN [72]	73.81%	89.85%	89.75%	69.51%	26.41%	34.67%
MAN	74.29%	**91.35%**	**91.51%**	71.16%	36.92%	44.00%
MCAN	**79.52%**	88.35%	87.70%	**73.52%**	**40.51%**	**48.22%**

network. SlowOnly [43] network is used to extract optic flow features. The batch-size is set to 24 for both source samples and target samples, resulting in a total of 48 samples fed to the model in each iteration. Since applying class weight calibration also changes the network convergence dynamics, grid searches are conducted to obtain the optimal training epochs on each experiment setting. More specifically, the network is trained for 20 epochs on **U-45→M-18** and **A-10→H-5**, 30 epochs on **H-14→U-7**, **M-45→U-18**, and 15 epochs on **H-10→A-5** and **U-14→H-7**. The learning rate is globally set to 0.001. $\alpha = 1$ for all settings. $\beta = 0.5$ for UCF-HMDB$_{partial}$ and MiniKinetics-UCF. $\beta = 2$ for HMDB-ARID$_{partial}$. $a = 5$ and $b = 1$ in Eq.3.10. For universality, while some other optimal values may produce even higher results, K is set to 4 for **U-14→H-7** and both settings in HMDB-ARID$_{partial}$, and 5 for others.

3.4.2 Results and Comparisons

The network performances on the target domains are compared with the previous PVDA and VUDA methods. Specifically, image-based methods including DANN [8], PADA [71], ETN [75], MK-MMD [54], MCD [81], and MDD [50] are adapted for transferring video features. TA^{3}N is reproduced based on their provided code and features [10]. Due to different experiment settings, PATAN [72] is reproduced. In general, results in Table 3.1 show that MAN and MCAN achieve results that surpass the previous highest results from PATAN or other methods. Based on the results of MAN, they show that the utilization of multi-modal features and multi-scale temporal pooling is indeed beneficial to enhancing PVDA performance.

More importantly, this enables more robust and less shifted class weight for outlier classes filtration. As a result, a relative improvement of 1.6% is achieved compared with PATAN [72] on UCF-HMDB$_{partial}$ and MiniKinetics-UCF. Particularly, MAN improves the results on **H-10→A-5** and **A-5→H-10** from 26.67%, 34.67% to 36.92% and 44.00%, respectively. Such significant performance uplift is largely due to the optic flow being much more robust to steep luminance changes between datasets. Moreover, it is worth noting that class weight additionally amplifies the benefits brought by multi-modal features since it can guide the training process much more precisely.

Furthermore, based on results from MCAN, they imply that the suppression of label distribution negative transfer is effective. This is supported by a 7.41% relative improvement on four datasets that has larger domain shifts, e.g **U-14→H-7**, **U-45→M-18**, **H-10→A-5**, and **A-10→H-5**. In conclusion, the significant improvements brought by our class calibration method prove that exploiting cluster structures of video features is a novel and valid idea. Admittedly, it is observed that class weight calibration is not effective on **H-14→U-7** and **M-45→U-18**. It is believed that this is largely due to the domain shifts being much smaller than other benchmarks, making the cluster structure being exploited less compatible with these settings. Nevertheless, even source-only networks, e.g. TRN [4] can achieve high PVDA performance in these settings.

3.4.3 Empirical Analysis

t-SNE visualization. To support our claims in section 3.3.3 about the cluster structures of video features, t-SNE [5] plots in Fig.3.4 for target features from **U-45→M-18** and **U-14→H-7** produced by MAN is obtained. Before any analysis, one should note that t-SNE plots are not a direct visualization of k-means clustering observations. Based on plots (b), (c), and (d), it can be observed that those incorrect predictions tend to be located near the centroids, while the entropy-based weighting blindly up-weighs the predictions of these features. This is in line with our motivation to implement the Eq.3.9, which is to down-weigh incorrect predictions. For **U-14→H-7**, the t-SNE plots show that the 2nd class and 6th class are weighted similarly by both calibration methods, but the entropy-based

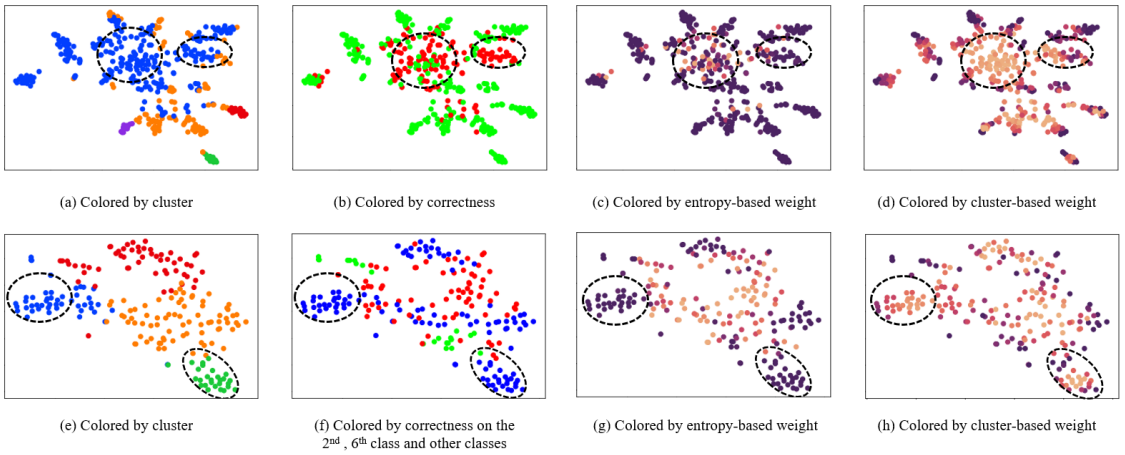

FIGURE 3.3: t-SNE [5] plots colored by different criteria. Plots (a) to (h) are t-SNE plots for features generated by MAN at the 5^{th} epoch. The usage of early epochs is intentional since the first few class weights are crucial to successful training. The first row and second row correspond to features of test samples from **U-45→M-18** and **U-14→H-7**. (a) and (e) is colored by cluster assignments. Correct predictions are colored green, and incorrect predictions are colored red in (b). Incorrect predictions are colored by red, 2^{nd} and 6^{th} classes are marked as green, and others are colored as blue in (f). H_{cls} in Eq.3.10 is removed to produce (c) and (g), and H_{ent} is removed to produce (d) and (h). The degree of weighing is shown by color in (c), (g), (d), (h), and they are directly comparable. Deeper colors mean the weight is higher. Black dotted circles in the first row mark the case where the entropy-based calibration can accidentally up-weigh incorrect predictions. The circles in the second row mark the case where the entropy-based calibration can overly up-weigh well-classified classes. If inconsistencies occur between observations based on k-means and the t-SNE plots, it is recommended to refer back to previous observations made in section 3.3.3.

weighing can produce a class weight with a high vector mean due to the over up-weighing of classes marked by dotted black circles. This is also in line with the motivation of designing Eq.3.9, which is to weigh predictions in a balanced way, i.e. referring to the relative position of each feature in its assigned cluster. Overall, t-SNE plots offer crucial visualizations to support the formulation of our weighting scheme, and they also suggest that the exploited cluster structures are not unique to a single experiment setting.

Class weight visualization. To understand the efficacy of calibrating the class weight and demonstrate that our calibration method is also applicable to other frameworks, an experiment on **U-45→M-18** is conducted using TRN based on the TSN backbone [2] equipped with [4] PADA [71] with and without the calibration. In general, TRN+PADA+Calibration achieves an accuracy of 69.50%, while

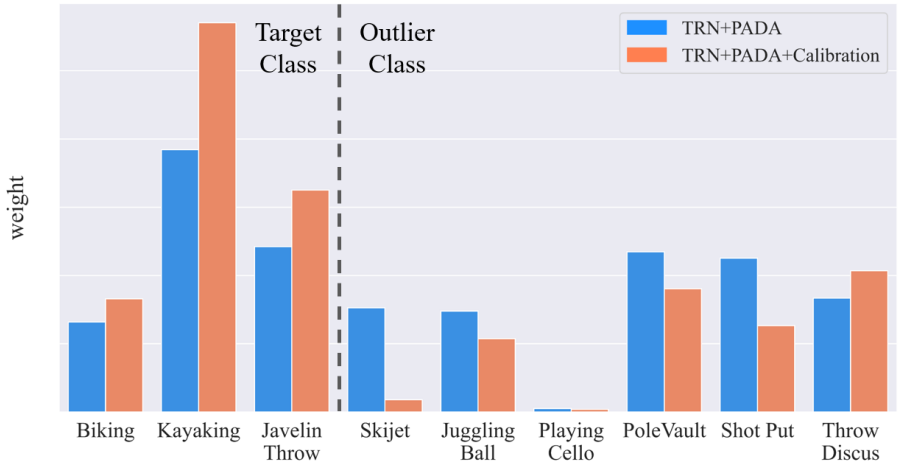

FIGURE 3.4: Class weight of each class in **U45→M18** at epoch 20 using TRN+PADA and TRN+PADA+class weight calibration. Classes from *Biking* to *Javelin Throw* are target classes, while others are outlier classes. The weights are all obtained as raw weights. A higher weight for target classes is better, and a lower weight for outlier classes is better. Best viewed in color.

TRN+PADA achieves 63.35%. For simplicity, only 9 classes in all 45 classes are shown in Fig.3.4. Based on the case of *Kayaking* and *Skijet*, our calibration method indeed suppresses the label distribution negative transfer because the weight for *Skijet* is significantly reduced. This also leads to the up-weighing of *Kayaking* since these two are similar. For other outlier classes that share less similarity with target classes, it is also observed that these outlier classes, e.g. *Juggling Ball*, *Pole Vault*, *Shot Put* are constantly down-weighted, and the target class, *Biking*, is up-weighted. For classes that share excessive similarities, e.g. *Javelin Throw* and *Throw Discus*, the up-weighing of *Javelin Throw* is relatively more than that of *Throw Discus*. This suggests that even if the network only has a limited ability to distinguish them, the class weight calibration can still promote the target classes and facilitate the positive transfer of relevant source data. Overall, results in Fig.3.4 show that our goal of promoting the positive transfer of relevant source data and suppressing the negative transfer of irrelevant source data is achieved. Moreover, this indicates that our calibration method could indeed work with other PVDA methods.

3.4.4 Ablation Study

Network ablation. To further analyze the efficacy of the proposed MCAN, perform ablation studies are performed to evaluate MCAN against its variants: (1)

MCAN w/o class weight is a variant that does not contain class weight updating over the entire training process, i.e. it is equivalent to MAN; (2) *MCAN w/o calibration* is a variant where class weight calibration is removed; (3) *MCAN w/o adversarial* is a variant without adversarial network, and (4) *MCAN w/o flow* is a variant where the optic flow estimation and extraction pipeline are removed. This thesis evaluates these variants on the challenging HMDB-ARID$_{partial}$ dataset to demonstrate the efficacy of MCAN, and the results are shown in Table 3.2.

TABLE 3.2: Ablation Studies on HMDB-ARID$_{partial}$

method	**H-10→A-5**	**A-10→H-5**
MCAN	**40.51%**	**48.22%**
MCAN w/o class weight	37.95%	37.33%
MCAN w/o calibration	36.92%	44.00%
MCAN w/o adversarial	35.90%	43.33%
MCAN w/o flow	23.67%	35.33%

Specifically, the first variant is meant to demonstrate the effectiveness of utilizing class weight, and the results indeed support that class weight can benefit PVDA performance. The second variant is set to justify the efficacy of class weight calibration. Compared with MCAN, the results imply that the calibration surely offers a positive performance uplift. The third variant is meant to demonstrate the importance of generating transferable features via domain adversarial networks, and the results also align with our intention. Last, the fourth variant demonstrates the dramatic performance uplift should thank the application of multi-modal features, e.g. optic flow. Moreover, while the calibration may not work efficiently if the predictions are too noisy, e.g. the bottom column of **H-10→A-5**, it is indeed effective on **A-10→H-5**.

K value ablation. In many methods that utilize k-means clustering, selecting an optimal K value through grid searches is essential. However, in this thesis, it is empirically demonstrated that the performance uplift can be achieved with a relatively wide range of K. To support this claim, multiple experiments are conducted on **U-14→H-7** by gradually increasing the K, and the results are shown in Fig.3.5. Compared with the baseline result from MAN, MCAN can achieve an average of 4.05% relative improvement when $K \in [1, 7]$. On the other hand, the performance quickly drops after K gets larger than 9, and the k-means clustering would not converge when K is greater than 11. Such a phenomenon is expected as

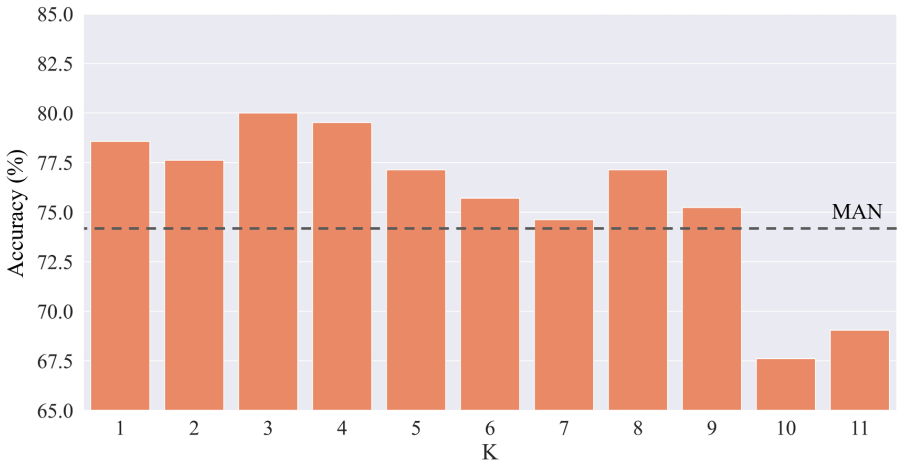

FIGURE 3.5: Ablation study of the K value on **U-14→H-7** using MCAN. The dotted line marks the baseline result from MAN.

our class weight calibration is built upon observations made for smaller K, and the features are assumed to be less separable. In conclusion, this section empirically demonstrates that our class weight calibration method is a practical method that is not excessively sensitive to the K value.

3.5 Conclusion

In this work, a novel multi-modality network MAN to improve video feature robustness across different domains significantly is proposed. Unlike previous works that only rely on RGB-only features, MAN explicitly represents motions via optic flow modality and utilizes multi-scale temporal pooling to enhance the predicting performance and class weight. To particularly aid the issue of label distribution negative transfer, it is proposed to calibrate the class weight of MAN, which brings us MCAN. The class weight calibration method in MCAN exploits cluster structures of video features to correctly promote positive transfers of relevant source data and suppress negative transfers of irrelevant source data simultaneously. The state-of-the-art PVDA performance of our networks are well justified by our extensive experiments across different PVDA benchmarks and subsequent analysis.

Chapter 4

Continuous Video Domain Adaptation

Chapter 3 delved into the problem of partial video domain adaptation, exploring ways to improve the performance of video action recognition models when faced with redundant classes in the source label space. While substantial progress was made, the proposed method did not fully attend to real-world scenarios where continuous adaptation is needed. In real-world applications, the data distribution often changes over time, which is a distinct challenge that our previous study did not address. Given this, Chapter 4 investigates to tackle the problem of continuous video domain adaptation. The aim is to extend the application range of video action recognition models, enabling them to adapt to evolving environments and maintain high performance over time. It is important to note that the work in this chapter does not utilize the partial assumption about the label space because this research intends to study fundamental problems such as continuous domain adaptation for video action recognition. Nonetheless, this does not exclude the possibility of integrating the insights and techniques from the two chapters in the future. By combining the strategies for dealing with partial and continuous domain adaptations, we may be able to create more robust and adaptable video action recognition methods.

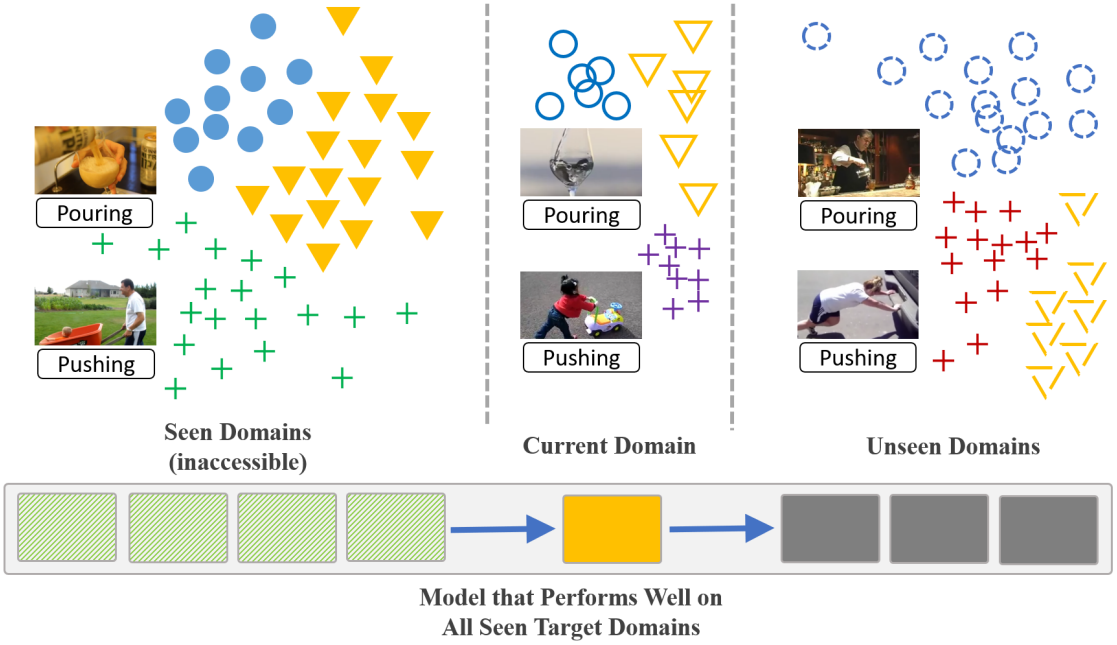

FIGURE 4.1: Illustration of how changing target domains are encountered in CVDA. The source model is adapted to each individual arriving domain, and seen domains are not accessible. In such a scenario, VUDA methods can forget previously learned knowledge, and it can be limited by storage issues. Therefore, it is nontrivial to tackle the continuous learning challenge in CVDA, and CART is proposed to address these challenges.

4.1 Introduction

Video action recognition using deep learning [2, 6, 45, 98] is a long-studied topic and has many applications [99–103]. Still, in real-world scenarios, model adaptation is often necessary due to domain shifts between the source and target domains. Therefore, Video-based Unsupervised Domain Adaptation (VUDA) methods [10, 104–106] were proposed to transfer models from a pre-training (source) domain to a static new (target) domain.

However, continuous adaptation is more desirable for real-world machine perception systems because they are running in continually changing environments, and the target domain distribution can change over time. Intuitively, assuming the source data can be stored and accessed, VUDA methods can continuously adapt to new arriving target domains. Still, this is sub-optimal as VUDA methods do not consider preserving knowledge learned on previous domains, making repeated adaptation necessary when seen domains are re-appearing. On the other hand, there is a strong motivation to enable continuous adaptation without any previous

Setups	Source Data	Target Supervision	Target Task	Data Type
UDA [8, 50, 54, 81]	Yes	No	Static Domain	Image
VUDA [10, 104–106]	Yes	No	Static Domain	Video
SFDA [77, 107, 108]	No	No	Static Domain	Image
SFVDA [110]	No	No	Static Domain	Video
CiDA [111]	No	No	Static Class-Incremental	Image
CL [112–114]	N/A	Yes	Continuous Class-Incremental	Image/Video
CDA [109]	N/A	Yes	Continuous Changing Domain	Image
Ours (CVDA)	**No**	**No**	**Continuous Changing Domain**	**Video**

TABLE 4.1: Differences between CVDA and previous similar setups

data, including the source data, because storing the massive source data and the accumulated target data for VUDA can be infeasible on platforms such as robots. For example, a public safety robot can encounter videos shot in changing environments (domains), e.g., day and night, sunny and foggy weather, even at the same location, and its onboard storage space can be very limited. To this end, the aforementioned scenario can be summarized as Continuous Video Domain Adaptation (CVDA), and the objective for CVDA methods is to obtain a video model that performs well on all seen domains. Specifically, to enable video models to continuously adapt to encountering target domains with only the unsupervised target data from the encountering domain, two goals need to be achieved: 1) the method should enable the model to learn helpful information from each encountering unsupervised target domain without source and previous target domain data, and 2) the method should retain the model performance on all seen target domains.

While VUDA methods are failing to achieve those two goals simultaneously, a common strategy to enable unsupervised learning without source data is pseudo labelling [77, 107, 108]. These methods focus on adapting to a static target domain and do not consider the accumulation of prediction errors in long-term continual adaptation scenarios. Over time, the accumulated error can significantly mislead the model and cause catastrophic forgetting. Meanwhile, works such as CDA [109] and most Continual Learning (CL) methods consider mitigating catastrophic forgetting and assume the target supervision exists. This is sub-optimal for videos as obtaining target supervision is time and labor-intensive. Limitations of existing works are listed in Table 4.1. Overall, without effective measurements to simultaneously tame the accumulation of prediction errors over time and learn from unsupervised target domains without source data, previous methods are generally incapable of performing well in the CVDA scenario.

Based on those two aforementioned goals, it is identified that the CVDA scenario

can be tackled by: *1) robust source-free learning of unsupervised changing target domains, and 2) mitigation of accumulated prediction errors that mislead models and cause catastrophic forgetting.* In view of this, this thesis proposes a **Confidence-Attentive network with geneRalization enhanced self-knowledge disTillation (CART)**. Firstly, CART uses target predictions as pseudo labels and deploys prototypical classification to refine noisy pseudo labels. To reduce the unreliability of pseudo labels when the target domain changes continuously, a new learning strategy is proposed where samples are learned attentively based on their prediction confidence such that high confidence samples are learned to lead the model update while low confidence but correct samples can also contribute to the acquirement of new knowledge. Secondly, observing that the source model is often less biased and more plastic than target models trained with noisy pseudo labels, it is proposed to regularize the current model parameter to behave similarly to the source model. This is achieved by constructing a self-knowledge distillation process enhanced by data generalization. Specifically, the current model is first fed with generalized, i.e., strongly augmented, data and the source model is fed with weakly augmented data. Next, the current model is enforced to mimic the output of the source model to ensure the current model behaves similarly to the source model. This ensures that the current model behavior is consistent with the less biased source model in a larger domain, i.e., the generalized target domain, and the accumulation of prediction errors is better reduced. Extensive experiments show that CART can effectively reduce the accumulation of prediction errors and achieve state-of-the-art continuous adaptation results on new dedicated CVDA benchmarks.

In summary, this chapter has made the following contributions: 1) to the best of our knowledge, this chapter of the thesis is the first research to touch on the concept of CVDA, where deep video models are required to continuously adapt to new target domains while retaining learned knowledge on seen domains without the source data and target supervision; 2) the challenges underlying CVDA are analyzed and CART is introduced to address these challenges by attentive learning of pseudo labels and generalization enhanced self-knowledge distillation; and 3) two dedicated CVDA benchmarks are constructed, namely Sports-DA$_{Conti.}$ and Daily-DA$_{Conti.}$. Extensive experiments demonstrate that CART achieves an average of 8.42% relative improvement in adaptation performances over previous state-of-the-art methods. Meanwhile, CART also has an average forgetting rate of -1.09%,

indicating that CART enables knowledge accumulation instead of forgetting.

4.2 Related Works

Video-based Unsupervised Domain Adaptation Unsupervised Domain Adaptation (UDA) methods aim to transfer source domain image knowledge to an unsupervised target image domain. The mainstream idea is to mitigate domain discrepancy [8, 50, 54, 81]. Recent UDA studies propose other ideas, such as varying the input space [115] or leveraging self-training [116]. In contrast, recent Video-based Unsupervised Domain Adaptation (VUDA) methods [10, 104–106, 110] focus on enabling efficient transfer of video models. A detailed introduction of those related works can be viewed in Section 2.2.1.

Continual Learning The main challenge in Continual Learning (CL) is to prevent catastrophic forgetting when learning new tasks [117]. Continual learning methods can be divided into two categories: replay-based [118] and regularization-based [112, 119, 120]. Replay-based methods memorize a certain amount of source data to prevent forgetting, while regularization-based methods [112, 119, 120] seek to regularize the model update without any source data to mitigate forgetting. Among many CL methods, Knowledge Distillation [121] (KD) is particularly effective and does not require accessing source data. It was originally used for knowledge transfer from big to small models and was considered an effective CL method [112, 119, 122]. In this research, it was found that it offers simple but effective regularization on the current model in terms of preventing the accumulation of prediction errors.

Continuous Domain Adaptation While most CL methods are developed for class-incremental scenarios, others focus on continuous domain adaptation. Early works such as CMA [123] and IADA [124] addressed the problem of adapting to evolving target image data. More recently, CDA [109] proposed the use of meta-learning to adapt to a series of supervised target domains. Studies such as CiDA [111] instead combine domain adaptation with class-incremental continual learning. In other pioneering works [125], the combined problem of test-time adaptation and continuous image domain adaptation was tackled.

Domain Generalization. Domain generalization is a well-studied problem in computer vision, and many related methods have been proposed [126–128]. Among these methods, data augmentation is particularly effective. Image-transform-based methods [129, 130] have shown that image transformations can generalize source data to out-of-distribution data, enabling better adaptation performance. More recent studies propose to leverage neural networks to manipulate the augmentation process of images [131, 132]. For a better understanding of domain generalization, readers are referred to related surveys [133]. In this research, it is shown that simple domain generalization techniques can further improve the effectiveness of self-knowledge distillation.

4.3 Continuous Video Domain Adaptation

4.3.1 Problem Definition

Given a pre-trained model consisting of a feature extractor g_0 with parameters θ_0 and classifier h_0 with parameters ϕ_0 trained on the source data $(\mathcal{X}^S, \mathcal{Y}^S)$, the goal is to adapt to a series of target domains $\mathcal{D}1, \mathcal{D}2, \mathcal{D}3, ...\mathcal{D}t, ...$ where $\mathcal{D}_t = \{x_{it}\}_{i=1}^{N_t}$ with Nt i.i.d. videos xit. The adaptation starts from the $t = 1$ moment, i.e., the source domain is not accessible, and $\mathcal{D}_t$ is only available at the current time step. The assumption is that the label spaces C_t across all domains are identical, and the source task $\mathcal{X}^0 \rightarrow \mathcal{Y}^0$ is the same as all target tasks $\mathcal{X}^t \rightarrow \mathcal{Y}^t$. In this manner, CVDA focuses on two metrics: 1) adaptation performance on $\mathcal{D}_t$ and 2) prediction performance on all seen target domains, e.g., $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_{t-1}$. By default, it is assumed that $t \geq 1$. Thus, without further notification, when t appears in a formula in this chapter, it refers to a time step greater than 0. For simplicity, i in x_{it} is omitted in some formulas, and x_t is weakly augmented by default.

The aforementioned CVDA setup is designed to address the growing demand for continuous adaptation to new video domains with limited data in many real-world applications. In CVDA, deep models cannot rely on source data and target supervision and are prone to noisy inputs. Moreover, CVDA also requires consistent performance on all previously seen target domains, indicating that methods designed for CVDA should preserve previous domain knowledge and further enable

the accumulation of knowledge across domains to improve the adaptation performance on the newly learned domain.

4.3.2 Methodology

Source Model Generation. Many existing Source-Free Domain Adaptation (SFDA) methods [77, 111, 134] require a specific source model preparation phase where modified training strategies, such as entropy loss [77, 135], are applied. In contrast, CART does not require specific source preparation. Moreover, besides proposing CVDA, this research is also a pioneering video domain adaptation work that is solely based on transformer-based networks. TimeSFormer [45] is used as the backbone for all involved methods in this chapter, for it is relatively lightweight, performative, and robust.

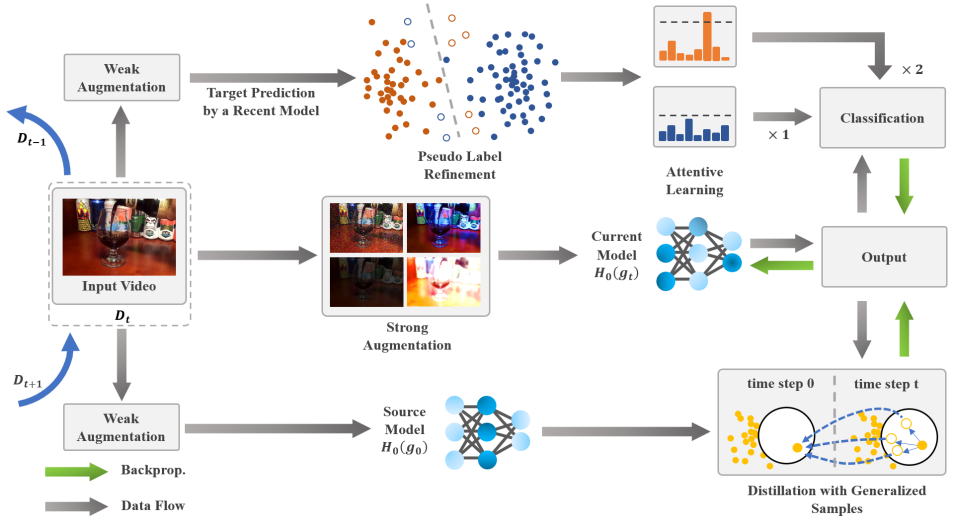

FIGURE 4.2: Graphical illustration of the proposed CART. Whenever a new domain $\mathcal{D}_t$ is met at time step t, CART first feeds weakly augmented samples to the source model $h_0(g_0)$ and a previous version of $h_0(g_t)$ saved recently. Next, CART feeds strongly augmented samples to the current model $h_0(g_t)$. With the model output, the classification loss $\mathcal{L}_{Acls}$ is obtained using Eqn. 4.5. Combining with the distillation loss $\mathcal{L}_{Adis}$, the loss to be optimized in CART is obtained, i.e., Eqn. 4.7. With the combined loss, With the combined loss, the current model $h_0(g_t)$ is updated with stochastic gradient descent until epochs on the data batch for $\mathcal{D}_t$ are completed.

Attentive Learning with Pseudo Labels. Modern deep classification models are commonly constructed as two parts, i.e., a feature extractor g_0 and a classifier h_0. UDA methods [8, 50, 54, 81] often aim to generate domain invariant features

by minimizing metrics that could reflect the degree of domain discrepancy. Thus, the loss to be optimized for a single source-target sample pair of samples in UDA methods can be summarized as follows:

$$\mathcal{L} = \mathcal{L}_{cls}(x_s, y_s) + \mathcal{L}_{dom}(x_s, x_t), \tag{4.1}$$

where $\mathcal{L}_{cls}$ is the source classification loss, $\mathcal{L}_{dom}$ is a domain loss that takes both source and target data as input, x_s is a source sample. This is not applicable in the proposed CVDA setup where source data are not accessible. In view of this, SFDA methods [77, 107, 108] propose either learning from pseudo labels or optimizing other types of unsupervised losses [77]. In this research, inspired by previous studies [77, 135], a simple yet effective way to obtain refined pseudo labels is taken. Formally, given target features $g_t(x_t)$ and target predictions $h_t(g(x_t))$, the centroids of target features can be obtained by the weighted average of all features on each class as follows:

$$c_k = \frac{\sum_{x_t \in \mathcal{X}^t} \delta_k(h_0(g_t(x_t)))g_t(x_t)}{\sum_{x_t \in \mathcal{X}^t} \delta_k(h_0(g_t(x_t)))}, \tag{4.2}$$

where δ_k is a softmax function and k indexes the k-th class. Then, by classifying each feature to the corresponding class of its closest centroid, the refined pseudo label is:

$$\hat{y}_t = \arg\min_k Dist(g_t(x_t), c_k), \tag{4.3}$$

where $Dist$ measures the cosine distance between a target feature to all centroids, notice that the parameter ϕ_0 of the source classifier h_0 is not optimized as previous works suggest that this can be unnecessary [77].

Refined pseudo labels can still contain a considerable amount of prediction errors when the domain shift is large. In such scenarios, even state-of-the-art refinement strategies can only provide minimal improvement, making regularizing the model to reduce the accumulation of prediction errors necessary. In view of this, it is proposed to learn from samples attentively. Specifically, each prediction is learned attentively based on their prediction confidence. Firstly, the prediction confidence of sample x_t is calculated as follows:

$$Conf(x_t) = max(\delta(h_0(g_t(x_t)))). \tag{4.4}$$

While previous methods [136] forsake all low-confidence examples, it was found that this is wasteful as there can be many low-confidence correct predictions in CVDA. Therefore, it is proposed to pay attention to high-confidence and low-confidence samples differently such that the former can lead the optimization while new knowledge can also be learned from the latter, i.e.,

$$\mathcal{L}_{cls} = \frac{1}{N_t}\sum_{i=1}^{N_t}\sum_{m=1}^{M}(\mathbb{1}(Conf(x_{it}) > \tau_m)\mathcal{L}_{ce}(x_{it}, \hat{y_{it}})), \tag{4.5}$$

where $\mathbb{1}(Conf(x_{it}) > \tau_m)$ masks out predictions whose confidence are lower than τ_m, and $\mathcal{L}_{ce}$ is a standard cross-entropy loss. In practice, it is found that setting $M = 2, \tau_1 = 0.1, \tau_2 = 0.5$ is generally sufficient.

Self-Knowledge Distillation with Generalized Samples. Though learning attentively from refined pseudo labels is effective, it cannot guarantee that the accumulation of prediction errors is reduced. This is because pseudo labels inevitably contain errors, and the error rate is likely to be reinforced when continuously adapting for a long time. In addition, the domain shifts between video domains are often large and such shifts can make the error rate grow rapidly as the model is misled continuously by noisy pseudo labels. Eventually, without the help of source and target supervision, the model may not be able to recover from such a state by itself. In view of this, inspired by some CL methods [112, 119], this chapter proposes to regularize the model update such that $h_0(g_t)$ is enforced to behave similarly as $h_0(g_0)$ via self-knowledge distillation. By enforcing the current model $h_0(g_t)$ to behave similarly to the source model $h_0(g_0)$, the accumulation of errors can be kept at a lower level such that the model can retain learned knowledge. Formally, the self-distillation loss is written as follows:

$$\begin{aligned}\mathcal{L}_{dis} &= \frac{1}{N_t}\sum_{i=1}^{N_t}\mathrm{KL}(\delta(h_0(g_0(x_{it})))|\delta(h_0(g_t(x_{it})))) \\ &= \frac{1}{N_t}\sum_{i=1}^{N_t}\delta(h_0(g_0(x_{it})))\log\frac{\delta(h_0(g_0(x_{it})))}{\delta(h_0(g_t(x_{it})))},\end{aligned} \tag{4.6}$$

where the source model $h_0(g_0)$ is regarded as a teacher model, and the model at time step t is considered as a student model. By minimizing $\mathcal{L}_{dis}$, the current model $h_0(g_t)$ can receive knowledge from the less biased $h_0(g_0)$ and keep itself less biased.

Notably, the distillation loss that uses the model from $t-1$ moment as the teacher model $\mathcal{L}'_{dis}$ is denoted. This is a common practice in previous works [112], but it is unsuitable in CVDA since 1) frequently saving g_{t-1} is space inefficient, and 2) models trained with pseudo labels are generally biased by accumulated prediction errors.

With the aforementioned process, it is observed that the accumulation of prediction errors is reduced. Still, it is found the results are sometimes unsatisfactory when the model is required to learn new knowledge while having minimal forgetting rates. Intuitively, enlarging the trade-off for $\mathcal{L}_{dis}$ can ensure a better reduction of the accumulation of prediction errors. However, it is observed that large trade-offs often limit the model to learn new knowledge. To achieve stronger regularization without damaging model plasticity, i.e., the ability to learn new knowledge, this chapter proposes to enhance the regularization on $h_0(g_t)$ by first strongly augmenting samples fed to it and then computing the distillation loss between the output of strongly augmented samples from $h_0(g_t)$ and the output of weakly augmented samples from $h_0(g_0)$. To this end, the overall training and evaluation pipeline is also illustrated in Algorithm 1, and Eqn. 4.5 and Eqn. 4.6 can be re-written to obtain the loss that is optimized in CART:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{Acls} + \mathcal{L}_{Adis} \\ &= \frac{1}{N_t}\sum_{i=1}^{N_t}\sum_{m=1}^{M}(\mathbb{1}(Conf(\mathcal{A}(x_{it})) > \tau_m)\mathcal{L}_{ce}(\mathcal{A}(x_{it}), \hat{y_{it}})) \\ &+ \frac{\alpha}{N_t}\sum_{i=1}^{N_t}\delta(h_0(g_0(x_{it})))\log\frac{\delta(h_0(g_0(x_{it})))}{\delta(h_0(g_t(\mathcal{A}(x_{it}))))},\end{aligned} \tag{4.7}$$

where $\mathcal{A}$ represents the strong augmentation applied to the input data x_{it}, $\mathcal{L}_{Acls}$ and $\mathcal{L}_{Adis}$ are classification and distillation loss that receive strongly augmented samples as their input from $h_0(g_t)$. The key motivation behind constructing $\mathcal{L}_{Adis}$ in Eqn. 4.7 is that, instead of only regularizing the model to respond similarly when the input is x_t, the model is regularized to respond similarly to the source model output $h_0(g_0(x_t))$ when the generalized version of the input, e.g., $\mathcal{A}(x_t)$, arrives. This enforces the model to act similarly to the source model not only in the target domain $\mathcal{D}_t$ but also on a generalized target domain $\hat{\mathcal{D}}_t$. Empirically, it is found that this strategy is more effective than distilling with $\mathcal{L}_{dis}$, indicating that generalization can allow better model plasticity while retaining previous knowledge.

Algorithm 1: The training pipeline of CART

Data: Streams of target data x_t on a target domain sequence $\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, ...\}$, Source model $h_0(g_0)$.
Result: Model $h_0(g_t)$ adapted to $\mathcal{D}_t$
for *Epochs on* $\mathcal{D}_t$ *in* $\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, ...\}$ **do**
  Weakly augment x_t;
  Obtain all source model outputs $h_0(g_0(x_t))$;
  Obtain all $\hat{y}_t$ via Eqn. 4.3 every few epochs;
  for x_t *in* $\mathcal{D}_t$ **do**
    Strongly Augment x_t as $A(x_t)$;
    Obtain classification and distillation loss via Eqn. 4.7;
    Back propagation;
  end
  Update parameter $\theta_{t-1} \rightarrow \theta_t$;
  Evaluation on $\mathcal{D}_t, \mathcal{D}_{t-1}, ..., \mathcal{D}_1$;
end

Notably, strong augmentation is also applied to $\mathcal{L}_{cls}$, i.e., $\mathcal{L}_{Acls}$, forming a structure similar to some previous works [136]. Such structure for classification is empirically proven to be effective by [136] in terms of improving adaptation performance, and it also simplifies the data processing pipeline of CART shown in Fig. 4.2.

4.4 Experiments

Experiments are conducted on new benchmarks designed for Continuous Video Domain Adaptation (CVDA). This method is compared against other representative methods and demonstrated to efficiently tackle the challenge of CVDA. In-depth analyses of modules in CART are presented as well.

4.4.1 Experimental Settings

Our experiments are conducted on two CVDA benchmarks: Daily-DA$_{Conti.}$, and Sports-DA$_{Conti.}$.

Daily-DA$_{Conti.}$ is modified based on Daily-DA proposed by MSVDA [137], it is constructed upon four mainstream datasets, e.g. Kinetics-600(K600) [138], HMDB51(HMDB) [23], ARID [93], and Moments-in-Time(MIT) [69]. Kinetics-600, HMDB51, and Moments-in-Time are three prevailing large datasets used

for video action recognition benchmarking. ARID is a dataset specifically built for action recognition in dark environments, and its data distribution drastically differs from others, making it challenging for video models to perform well on it. 11 classes are collected in this benchmark, and more than 20000 videos are available. Given that Daily-DA is a multi-source domain adaptation benchmark, the following modifications were made to enable it as a benchmark for Continuous Video Domain Adaptation (CVDA): 1) Kinetics-600 is regarded as a fixed source domain, and 2) others are split into two non-overlapping parts equally. The dataset was split with the intention of training methods on a prolonged sequence of domains so that the benchmark can better simulate real-world scenarios where changing target domains arrive continuously and sometimes previously seen domains can re-appear. As a result, HMDB51 is split into HMDB_1 and HMDB_2. Splits for Kinetics600 and MIT also follow the same naming convention. With training data prepared, the main experiments were performed on two sequences, e.g., $\text{ARID}_1 \rightarrow \text{MIT}_1 \rightarrow \text{HMDB}_1 \rightarrow \text{ARID}_2 \rightarrow \text{MIT}_2 \rightarrow \text{HMDB}_2$ and $\text{HMDB}_1 \rightarrow \text{ARID}_1 \rightarrow \text{MIT}_1 \rightarrow \text{HMDB}_2 \rightarrow \text{ARID}_2 \rightarrow \text{MIT}_2$. For simplicity, these two sequences will be quoted in following paragraphs as ARID→MIT→HMDB51 and HMDB51→MIT→ARID, respectively in following tables, figures, and paragraphs.

Sports-DA$_{Conti.}$ is processed similarly to Daily-DA. It is also an adapted version of the original Sports-DA proposed by MSVDA [137]. This benchmark contains data from three mainstream datasets, e.g., Sports-1M [68], Kinetics-600 [138], and UCF101 [24]. 23 classes were involved, and over 40000 videos are available for benchmarking. Similarly, UCF101 is considered as the fixed source dataset, and the training data of Kinetics-600 and Sports-1M is split into two non-overlapping splits equally to create four changing batches of data to adapt to., e.g., K600_1, K600_2, Sports1M_1, and Sports1M_2. The sequence used during experimentation is $\text{Sports1M}_1 \rightarrow \text{K600}_1 \rightarrow \text{Sports1M}_2 \rightarrow \text{K600}_2$. For simplicity, this sequence is referred to as Sports1M→Kinetics600. Specific dataset information for both Daily-DA$_{Conti.}$ and Sports-DA$_{Conti.}$ is displayed in Table 4.2.

Statistics	Daily-DA$_{Conti.}$	Sports-DA$_{Conti.}$
Video Classes #	11	23
Training Video #	A:3,792 / H:770 / M:5,500 / K600:10,639	U:2,145 / S:14,754 / K600:19,104
Testing Video #	A:1,768 / H:330 / M:550 / K:725	U:851 / S:1,900 / K:1,961

TABLE 4.2: Statics of Daily-DA$_{\text{Conti.}}$ and Sports-DA$_{\text{Conti.}}$ A, H, M, and K600 refer to ARID, HMDB51, Moments-in-Time, and Kinetics-600. Notice that these two datasets use a different subset of Kinetics-600.

Class ID	ARID Class	HMDB51 Class	Moments-in-Time Class	Kinetics-600 Class
0	Drink	drink	drinking	drinking shots
1	Jump	jump	jumping	jumping bicycle, jumping into pool, jumping jacks
2	Pick	pick	picking	picking fruit
3	Pour	pour	pouring	pouring beer
4	Push	push	pushing	pushing car, pushing cart, pushing wheelbarrow, pushing wheelchair
5	Run	run	running	running on treadmill
6	Walk	walk	walking	walking the dog, walking through snow
7	Wave	wave	waving	waving hand
8	Sit	sit	sitting	falling off chair
9	Stand	stand	standing	snatch weight lifting
10	Turn	turn	turning	pirouetting

TABLE 4.3: List of action classes for Daily-DA$_{\text{Conti.}}$.

Class ID	UCF101 Class	Sports-1M Class	Kinetics-600 Class
0	Archery	archery	archery
1	Baseball Pitch	baseball	catching or throwing baseball, hitting baseball
2	Basketball Shooting	basketball	playing basketball, shooting basketball
3	Biking	bicycle	riding a bike
4	Bowling	bowling	bowling
5	Breaststroke	breaststroke	swimming breast stroke
6	Diving	diving	springboard diving
7	Fencing	fencing	fencing (sport)
8	Field Hockey Penalty	field hockey	playing field hockey
9	Floor Gymnastics	floor (gymnastics)	gymnastics tumbling
10	Golf Swing	golf	golf chipping, golf driving, golf putting
11	Horse Race	horse racing	riding or walking with horse
12	Kayaking	kayaking	canoeing or kayaking
13	Rock Climbing Indoor	rock climbing	rock climbing
14	Rope Climbing	rope climbing	climbing a rope
15	Skate Boarding	skateboarding	skateboarding
16	Skiing	skiing	skiing crosscountry, skiing mono
17	Sumo Wrestling	sumo	wrestling
18	Surfing	surfing	surfing water
19	Tai Chi	t'ai chi ch'uan	tai chi
20	Tennis Swing	tennis	playing tennis
21	Trampoline Jumping	trampolining	bouncing on trampoline
22	Volleyball Spiking	volleyball	playing volleyball

TABLE 4.4: List of action classes for Sports-DA$_{\text{Conti.}}$.

All methods are implemented using the PyTorch [95] library in this thesis. TimeSFormer [45] serves as the backbone. UDA/VUDA [8, 62, 81], SFDA [77], and TTA [125, 135] methods are involved as baselines. For UDA/VUDA methods, source data is made available to them, while SFDA and TTA methods are directly employed. Additionally, to ensure fair comparisons with other methods, state-of-the-art methods from related but different research works [8, 62, 77, 81, 125, 135] are adapted. SHOT [77] is directly employed without further modifications as it is compatible with CVDA settings. For UDA [8, 81], and VUDA [81] methods, source data is made available. Instead of feeding 16 samples per GPU, a total of 32 samples (16 sources and 16 targets) are fed to a GPU when UDA/VUDA methods are utilized. Since UDA/VUDA methods sometimes encounter stability issues [139] in CVDA, gradient clipping is applied to stabilize the training. For TTA methods [125, 135], the training and testing split is combined as one data stream. The

model is iterated directly on this data stream and will randomly encounter either a training or a testing sample, though only the accuracy of the testing samples is recorded. The source model is generated based on weights trained on Kinetics-400 [25]. The SGD optimizer is used with a default learning rate of 0.001. All involved methods are trained for 10 epochs on each domain before the next domain arrives. The learning rate is decreased to 0.0001 at epoch 5 for UDA/VUDA methods, SFDA methods, and CART. Pseudo labels are generated based on task predictions using a previous version of $h_0(g_t)$ that is saved every 5 epochs. A value of $\alpha = 5$ is set in Eqn.4.7 as it offers a reasonable level of regularization without damaging model plasticity. For the distillation process, the augmentation set used is inspired by RandAugment[140]. For weak augmentation, only some common augmentations such as center cropping and normalization [2, 6, 45] are leveraged. For strong augmentation, random HLS color variation, Gaussian noise, camera noise, flipping, etc. are included. The distillation temperature is set to 2 and the trade-off for distillation to 5. For the attentive learning process, $m = 2, \tau_1 = 0.1, \tau_2 = 0.5$ are set. Comprehensively, this means low-confidence samples are learned once while high-confidence samples are learned twice. For the backbone, TimeSFormer [45] is employed, which is an efficient and robust transformer-based network. This network achieves similar results to bigger models such as Swin [141] and ViViT [44]. Pre-trained weights based on the Kinetics-400 [25] dataset are loaded. For efficiency and accuracy, the first 4 blocks of the TimeSFormer are frozen. Training occurs on each data split for 10 epochs. The learning rate is set to 0.001 and decreases tenfold every 5 epochs. For pseudo label generation, CART re-generates the pseudo label every 5 epochs. Stochastic Gradient Descent (SGD)[142] optimizer is utilized with weight decay set to 0.0001 and momentum set to 0.9. Nesterov momentum[143] is also leveraged. A batch size of 96 with 16 samples per GPU is used. Each video clip contains 8 frames sampled from 8 segments of a video.

4.4.2 Results and Comparisons

In this study, the performance of CART is evaluated against several existing state-of-the-art methods, including SFDA methods such as SHOT [77], UDA/VUDA methods including DANN [8], MCD [81], and ACAN [62], and test-time adaptation methods TENT [135] and CoTTA [125]. Negative learning rates indicate the model is learning instead of forgetting knowledge. The average forgetting rate

is calculated between $h_0(g_t)$ and $h_0(g_{t-1})$ on all seen domains. Notably, average forgetting rates can be numerically small, but it can quickly lead to significant performance regression as time step t increases.

time →								
method	$ARID_1$	MIT_1	$HMDB_1$	$ARID_2$	MIT_2	$HMDB_2$	Mean Acc.	Mean Forget
Source-Only	20.81%	**26.91%**	40.00%	20.81%	26.91%	40.00%	29.24%	N/A
SHOT	21.55%	24.91%	29.39%	21.95%	25.27%	37.27%	26.72%	-1.59%
ACAN	21.67%	25.73%	38.03%	21.48%	24.85%	33.55%	27.55%	7.42%
DANN	**21.78%**	25.27%	37.88%	21.32%	26.00%	39.39%	28.61%	-0.81%
MCD	15.33%	23.45%	31.21%	12.10%	19.09%	25.45%	21.11%	1.24%
TENT	20.81%	26.18%	36.97%	20.76%	25.27%	36.36%	28.03%	2.78%
CoTTA	20.53%	**26.91%**	39.39%	18.27%	26.73%	36.67%	28.68%	0.66%
CART	**21.78%**	**26.91%**	**45.76%**	**26.19%**	**27.09%**	**47.58%**	**32.55%**	**-3.83%**

TABLE 4.5: Domain adaptation results on ARID→MIT→HMDB51

time →								
method	$HMDB_1$	$ARID_1$	MIT_1	$HMDB_2$	$ARID_2$	MIT_2	Mean Acc.	Mean Forget
Source-Only	40.00%	20.81%	26.91%	40.00%	20.81%	26.91%	29.24%	N/A
SHOT	43.94%	24.55%	25.64%	42.73%	23.42%	26.73%	31.17%	2.33%
ACAN	35.31%	20.84%	27.51%	37.25%	21.81%	25.62%	28.05%	4.05%
DANN	32.73%	20.53%	25.27%	35.15%	20.87%	24.91%	26.57%	**-0.51%**
MCD	31.82%	20.59%	22.73%	29.39%	15.84%	19.09%	23.24%	2.13%
TENT	36.36%	20.81%	25.82%	36.97%	20.81%	25.27%	25.27%	0.27%
CoTTA	37.88%	20.87%	26.36%	39.39%	20.70%	26.36%	28.59%	1.63%
CART	**46.06%**	**27.94%**	**28.73%**	**49.39%**	**29.19%**	**30.18%**	**35.24%**	-0.19%

TABLE 4.6: Domain adaptation results on HMDB51→MIT→ARID

Results in Table 4.5-4.7 show that the novel CART achieves state-of-the-art adaptation results consistently on three challenging benchmarks, outperforming the best-performing prior UDA/VUDA, SFDA, and TTA methods by an average improvement of 11.32% on HMDB51→MIT→ARID, 13.06% on HMDB51→ARID→MIT, and 0.89% on Sports1M→Kinetics600. It is identified that previous UDA/VUDA methods [8, 62, 81] are surpassed by the methods in this study by a large margin in terms of both average adaptation performances and sometimes the average forgetting rates, indicating that CART can remain stable even without the support of source data. Specifically, CART is obtaining the lowest average forgetting rate in Table 4.5 while being competitive in Table 4.6,4.7 when compared to UDA/VUDA methods that can access the source data, meaning that accumulation of prediction error is effectively mitigated even without the help of source data. With the low forgetting rate and high adaptation performance, the empirical results fully justify

time →						
method	Sports1M$_1$K600$_1$		Sports1M$_1$K600$_2$		Mean Acc.	Mean Forget
Source-Only	80.89%	89.91%	80.89%	89.91%	85.40%	N/A
SHOT	82.37%	93.32%	82.00%	93.43%	87.78%	1.06%
ACAN	81.21%	91.36%	82.44%	91.36%	86.59%	2.52%
DANN	80.16%	91.08%	81.37%	91.03%	85.91%	-0.34%
MCD	79.42%	88.43%	78.63%	85.58%	83.02%	0.51%
TENT	79.37%	90.27%	80.63%	90.52%	85.20%	**-0.54%**
CoTTA	79.63%	87.67%	75.47%	79.82%	80.39%	5.63%
CART	**82.53%**	**94.80%**	**82.58%**	**94.34%**	**88.56%**	0.73%

TABLE 4.7: Domain adaptation results on Sports1M→Kinetics600

that CART has effectively tackled the challenge brought by CVDA and enables the continuous learning of a global model.

Furthermore, it is observed that both SFDA methods [77] and UDA/VUDA [8, 62, 81] are achieving unsatisfactory results and sometimes worse than the source model. For SFDA methods, e.g., SHOT [77], it is observed that they are particularly prone to extremely noisy target domains, such as the ARID domain. Without measurements to reduce the accumulation of prediction errors, the continuous adaptation performance of SHOT[77] drops constantly and heavily. It is also found that, even with source data being accessible, UDA/VUDA methods are still incapable of adapting to continuously changing domains. Such results are believed to be caused by: 1) the Daily-DA$_{Conti.}$ is very challenging, and the domain adversarial network in UDA/VUDA can be unstable when the target domain is frequently varying. Finally, it is observed that TTA methods [125, 135] are also achieving degraded results. The belief is that this is because TTA methods cannot utilize samples effectively as they discard samples once learned, while their regularization [125, 135] may be too strong to allow learning of new knowledge.

4.4.3 Ablation Studies

To analyze the effectiveness of the two modules in CART, a series of ablation studies on the ARID→MIT→HMDB51 sequence are conducted and each module is carefully removed to investigate their effectiveness. Results are displayed in Table. 4.8. First, the effectiveness of enhancing distillation with data generalization by removing the generalization, i.e., optimize $\mathcal{L}Acls + \mathcal{L}dis$, is evaluated, i.e., optimize $\mathcal{L}_{Acls} + \mathcal{L}_{dis}$. Results suggest that the average adaptation performance dropped by 1.60%, and the average forgetting rate increased by 49.09%, indicating

method	Loss	Mean Acc.	Mean Forget
CART	$\mathcal{L}_{Acls} + \mathcal{L}_{Adis}$	**32.55%**	**-3.83%**
	$\mathcal{L}_{Acls} + \mathcal{L}_{dis}$	32.01%	-1.95%
	$\mathcal{L}_{cls} + \mathcal{L}_{dis}$	30.26%	-3.28%
	$\mathcal{L}_{cls}$	28.38%	0.60%
	$\mathcal{L}_{Acls} + \mathcal{L}_{Adis}, m = 1, \tau = 0$	32.16%	-3.48%
	$\mathcal{L}_{Acls} + \mathcal{L}_{dis}, \alpha = 10$	29.49%	1.19%
	$\mathcal{L}_{Acls} + \mathcal{L}'_{Adis}$	29.11%	2.57%

TABLE 4.8: Ablation study on ARID→MIT→HMDB51

the model is much more likely to forget previously learned knowledge. Next, strong augmentation is further removed entirely, i.e., optimize $\mathcal{L}_{cls} + \mathcal{L}_{dis}$, and find the adaptation performance is dropped heavily, indicating the accumulation error is less reduced and leading to degraded performance. Lastly, distillation in CART is completely removed, and it is found that the model performance decreases again. Overall, while distillation enables a significant reduction of accumulation of errors, the performance can receive another leap when using data generalization. On the other hand, the effectiveness of the proposed attentive learning of pseudo labels is also validated. Thus, $M = 1, \tau = 0$ is set to transform $\mathcal{L}Acls$ into a standard cross-entropy loss. Results reflect that our attentive learning strategy can indeed help to reduce the accumulation of prediction errors and improve model adaptation performance. Finally, to validate two claims made in Section 4.3.2 that distilling from the model at $t-1$ moment is sub-optimal and simply setting trade-off for distillation high can limit model plasticity, CART is optimized using $\mathcal{L}_{Acls} + \mathcal{L}_{dis}, \alpha = 10$ and $\mathcal{L}_{Acls} + \mathcal{L}'_{Adis}$, respectively. Results from both experiments have sufficiently proven our claims and justify the design of CART.

4.4.4 Result Analysis

Adaptation Performance Visualization. To understand how different methods react to continuous domain shifts, plot accuracy curves are plotted in Fig. 4.3. It also aims to visualize how the reduction of accumulated error is translated into an increase in adaptation performance. The experiment on ARID→MIT→HMDB51 is conducted and testing results on all domains is plotted even if the model has not seen them. According to Fig. 4.3, CART is achieving performance gains, particularly on the HMDB51 dataset, while the performances of some other methods are constantly dropping over time.

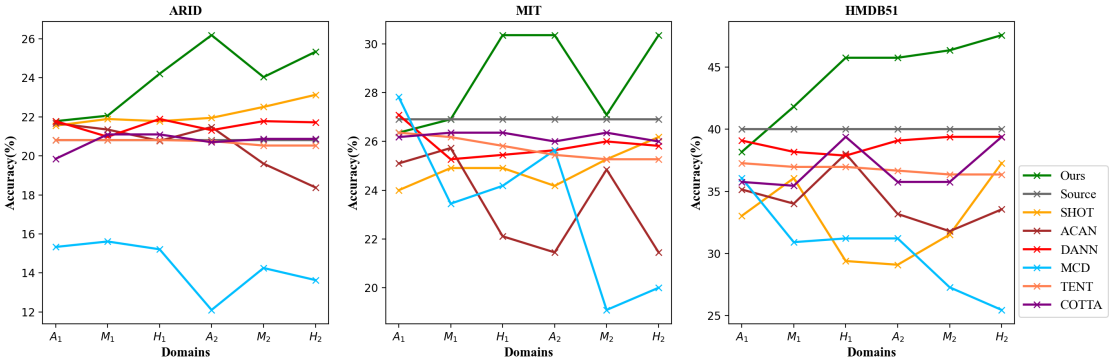

FIGURE 4.3: Accuracy curves for different methods on the ARID→MIT→HMDB51 sequence. A_1, M_1, H_1, A_2, M_2, and H_2 denote $ARID_1$, MIT_1, $HMDB_1$, $ARID_2$, MIT_2, and $HMDB_2$.

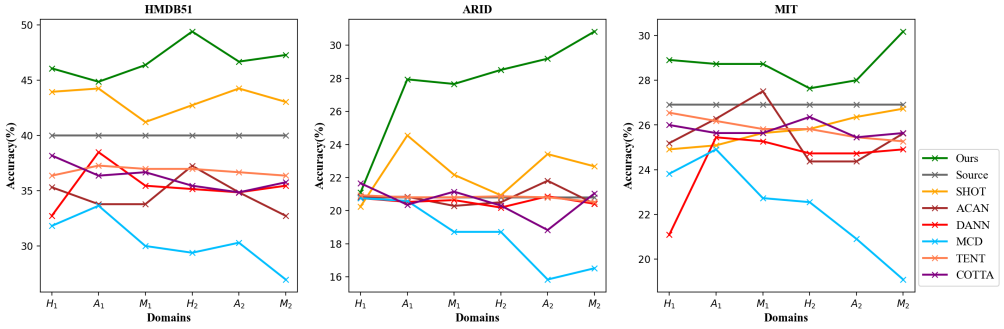

FIGURE 4.4: Accuracy Curves on HMDB51→ARID→MIT sequence

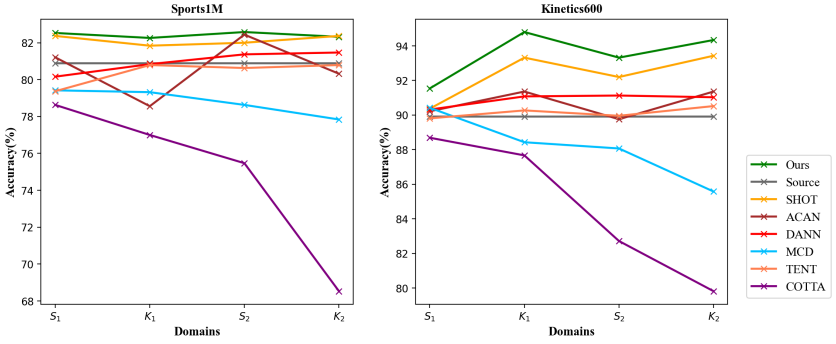

FIGURE 4.5: Accuracy Curves on Sports1M→Kinetics600 sequence

Specifically, it is found that most methods struggle to surpass the source model performance, indicating that the accumulated prediction errors are built up over time, and the model forgets learned knowledge. It is also found that ARID is particularly challenging as the accuracy of pseudo labels generated on this domain by the source model can be as low as 20%, making models more likely to be misled when adapting to this domain. Additionally, the curves for the ablation study are also plotted and it is shown that CART achieves the overall best performance while

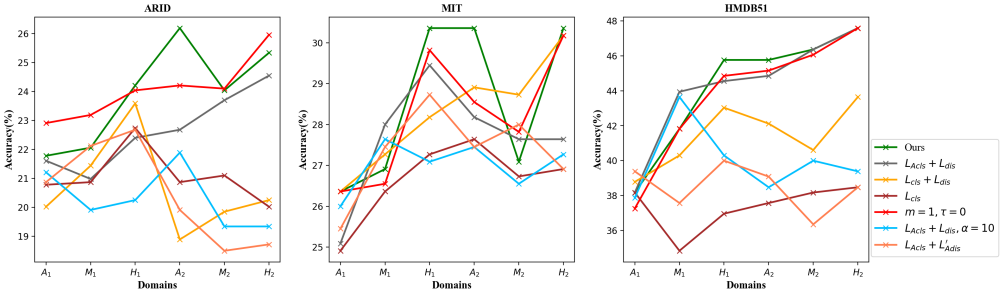

FIGURE 4.6: Accuracy Curves for Ablation Studies

other settings can have a certain level of performance regression. In summary, the results of CART on ARID prove that our method can learn from some of the most challenging target domains while also being performative on others, meaning that this research had effectively controlled the level of accumulated prediction errors in the learned model.

Self-Attention Visualization. Empirically, transformer-based networks rely on attending to a specific object in inputs to extract video features. Therefore, the attention map output from attention heads should be visualized as the time step t increases. Saliency maps are shown in Fig. 4.7 and Fig. 4.8. They show that the attention head gradually loses focus on specific objects, e.g., human arms, human

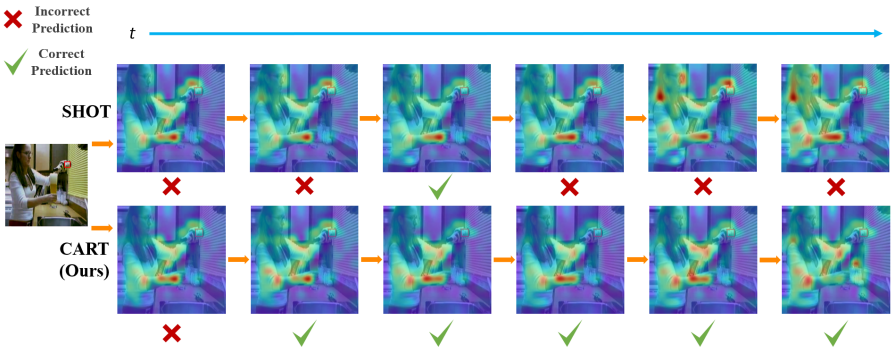

FIGURE 4.7: **Saliency maps that show SHOT can gradually mislead the model and result in incorrect predictions (marked by red crosses), while the proposed CART can maintain the focus of the model and enable the model to generate correct predictions (marked by green ticks)**. The shown attention maps are obtained from the first spatial attention head in the 10^{th} layer of the TimeSFormer network using the patch in the red box as a query. Models trained by SHOT and CART are evaluated on the same challenging testing sample, and the saliency maps for the same attention head are displayed above. Warmer colors, e.g., red and orange area, indicate the query is activating keys of those patches, while colder colors, e.g., green and purple, indicate those areas are less attended to by the attention head.

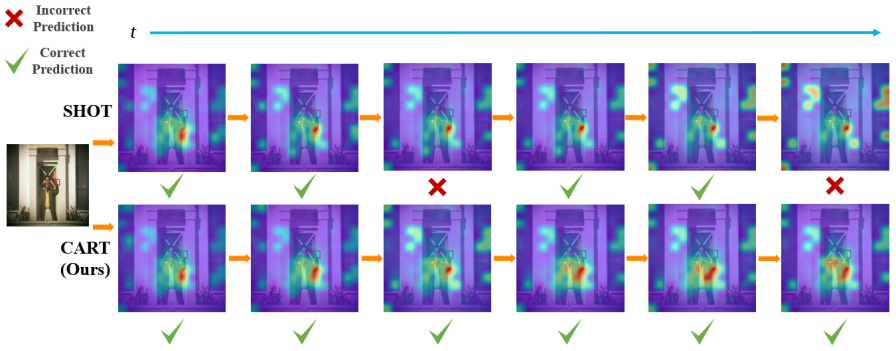

FIGURE 4.8: Saliency map visualization similar to Fig. 4.7. The action performed in the video is *'waving'*. The query is set to near the arm of the man (red bounding box). The results suggest that the model adapted by SHOT is shifting its focus onto irrelevant backgrounds while the model adapted by CART can attend to the human body more.

body, etc., when SHOT [77] is employed. In contrast, the saliency maps of CART demonstrate that the model consistently focuses on related objects and evolves to classify this challenging input correctly. In sum, Fig. 4.7 demonstrates that our novel regularization method can successfully control the detrimental drift of model parameters and thus improve the CVDA performance of modern video models.

4.5 Conclusion

This work proposes a novel Confidence-Attentive network with geneRalization enhanced self-knowledge disTillation (CART) to tackle the new Continuous Video Domain Adaptation (CVDA) task. In CVDA, it is assumed that target domains are continuously arriving and previously seen domains are unavailable. Without available source data, CART learns from refined pseudo labels attentively to avoid being misled by accumulated prediction errors. CART further reduces the accumulation of prediction errors by deploying a generalization enhanced self-knowledge distillation module. Extensive experiments justify that these two modules can effectively tackle challenges in CVDA, resulting in state-of-the-art adaptation results on dedicated CVDA benchmarks.

Chapter 5

Conclusion and Future Works

5.1 Conclusions

In this thesis, we present our work on the practical domain adaptation of video action recognition models. Conventional Video-based Unsupervised Domain Adaptation (VUDA) methods largely follow the standard assumption of Unsupervised Domain Adaptation (UDA), which assumes that the model is transferred from a single source domain to a single unsupervised target domain while key characteristics, such as label space and task, are identical. However, in real-world applications, this might not be the case. Therefore, for the first work in this thesis, we are inspired by the pioneering work in Partial Video Domain Adaptation (PVDA) where the source domain label spaces subsume the target label space and proposed a state-of-the-art adaptation method for this case, namely Multi-Modality Cluster-Calibrated Partial Adversarial Network (MCAN). In the second part of this thesis, we focus on enabling continuous adaptation of deep video models and first define the problem of Continuous Video Domain Adaptation (CVDA). This problem aims to solve the challenge that continuous domain adaptation of video models can lead to problems including catastrophic forgetting, rendering the model unable to perform equally well on learned domains.

More specifically, Chapter 3 proposes a novel multi-modality network MAN to improve video feature robustness across different domains significantly. Unlike previous works that only rely on RGB-only features, MAN explicitly represents

motions via optic flow modality and utilizes multi-scale temporal pooling to enhance the predicting performance and class weight. To particularly aid the issue of label distribution negative transfer, this research proposes to calibrate the class weight of MAN, which brings us MCAN. The class weight calibration method in MCAN exploits cluster structures of video features to correctly promote positive transfers of relevant source data and suppress negative transfers of irrelevant source data simultaneously. The state-of-the-art PVDA performance of our networks are well justified by our extensive experiments across different PVDA benchmarks and subsequent analysis.

Chapter 4 proposes a novel Confidence-Attentive network with geneRalization enhanced self-knowledge disTillation (CART) to tackle the new Continuous Video Domain Adaptation (CVDA) task. In CVDA, it is assumed that target domains are continuously arriving and previously seen domains are unavailable. Without available source data, CART learns from refined pseudo labels attentively to avoid being misled by accumulated prediction errors. CART further reduces the accumulation of prediction errors by deploying a generalization enhanced self-knowledge distillation module. Extensive experiments justify that these two modules can effectively tackle challenges in CVDA, resulting in state-of-the-art adaptation results on dedicated CVDA benchmarks.

In summary, this thesis focuses on expanding the scope of application of previous VUDA methods by investigating two specific domain adaptation problems, i.e., PVDA and CVDA. To effectively tackle each problem, each chapter identifies the key challenges of those problems and proposes novel solutions to address those challenges accordingly. In Chapter 3, our method improves over previous state-of-the-art methods by a large margin, and Chapter 4 defines novel scenarios and first conducts extensive pioneering experiments on the problem of CVDA. This thesis has achieved its planned goal while future works can keep integrating new methods that aid the problem of domain adaptation for video action recognition models.

5.2 Future Works

In total, two major works are presented in this thesis for practical video domain adaptation. Both works have achieved efficient improvements over previous methods and the success of those methods can lead us to some further in-depth research about practical video domain adaptation.

For MCAN proposed in Chapter 3, it is observed that integration of new modalities can effectively reduce the degree of bias of the target class estimation. This means that by integrating optic flow modality, the accuracy of the RGB stream could also potentially improve, given that the estimation of target classes becomes more precise. Therefore, it is pointed out that it is possible to integrate other kinds of modalities to further improve the efficiency of target class estimation. Some common practice is to also include inputs such as text and audio, while it is also possible to include pre-processed contour as input. On the other hand, since our calibration strategy is largely empirical, further understanding of the fundamental mechanism of such a method and this theis envision that further work can resolve the issue that the calibration method can bring performance regression when the adaptation performance is high, e.g., above 85%. For CART, it is clear that current source-free training methods are still largely limited due to low initial accuracy on the target datasets. While it is believed that multi-modality methods might be optional for such a dilemma, relevant research is still scant. On the other hand, while our self-distillation method is already achieving satisfying results, it is believed that it is possible to further boost the adaptation results as it is found that the domain knowledge from one target domain can also be utilized for the adaptation of other target domains. It is possible to identify such relations and set up regularization rules that focus more on such relationships.

To this end, it is identified that introducing partial domain adaptation to CVDA is also a valid research topic as it is natural that new arriving domains might not have identical labels as previous target domains and the source domain. In such cases, it will become challenging to identify target classes within the context of source-free and continuous domain adaptation. While this thesis can largely leverage the idea proposed in MCAN, the estimation of target classes needs to be treated with more caution as the accuracy of estimation is not guaranteed. Still, great challenges

come with great rewards. Such a combination can further enable video domain adaptation to be more practical and general.

List of Author's Awards, Patents, and Publications[1]

Conference Proceedings

- **Xiyu Wang***, Yuecong Xu*, Jianfei Yang, Kezhi Mao, 'Calibrating class weights with multi-modal information for partial video domain adaptation' in Proceedings of the 30th ACM International Conference on Multimedia, pp. 3945-3954. 2022.

[1]The superscript * indicates joint first authors

Bibliography

[1] Yann LeCun et al. Generalization and network design strategies. *Connectionism in perspective*, 19(143-155):18, 1989. xv, 10

[2] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36, Amsterdam, Netherlands, 2016. Springer. xv, 1, 11, 12, 30, 35, 38, 44, 56

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. xv, 14

[4] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 803–818, Munich, Germany, 2018. Springer. xv, 12, 24, 26, 29, 31, 32, 36, 37, 38

[5] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11):2579–2605, 2008. xvi, 37, 38

[6] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, Santiago, Chile, 2015. IEEE. 1, 12, 30, 44, 56

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, Honolulu, HI, USA, 2017. IEEE. 1, 13, 30

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2, 16, 18, 19, 26, 28, 29, 36, 45, 47, 49, 55, 56, 57, 58

[9] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006. 2, 17, 18

[10] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6321–6330, Long Beach, CA, USA, 2019. IEEE. 2, 19, 21, 23, 26, 28, 34, 36, 44, 45, 47

[11] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11815–11822, New York, NY, USA, 2020. AAAI. 2, 19, 23, 26, 28

[12] Chris Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988. 8

[13] Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64:107–123, 2005.

[14] Bhaskar Chakraborty, Michael B Holte, Thomas B Moeslund, and Jordi Gonzalez. Selective spatio-temporal interest points. *Computer Vision and Image Understanding*, 116(3):396–410, 2012. 8

[15] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 8

[16] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II 10*, pages 650–663. Springer, 2008. 8

[17] James W Davis and Aaron F Bobick. The representation and recognition of human movement using temporal templates. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 928–934. IEEE, 1997. 8

[18] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013. 9

[19] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part II 9*, pages 428–441. Springer, 2006. 9

[20] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Activity representation with motion hierarchies. *International journal of computer vision*, 107: 219–238, 2014. 9

[21] Saima Nazir, Muhammad Haroon Yousaf, and Sergio A Velastin. Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition. *Computers & Electrical Engineering*, 72:660–669, 2018. 9

[22] Xiaojiang Peng, Changqing Zou, Yu Qiao, and Qiang Peng. Action recognition with stacked fisher vectors. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 581–595. Springer, 2014. 9

[23] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563, Barcelona, Spain, 2011. IEEE. doi: 10.1109/ICCV.2011.6126543. 9, 19, 20, 21, 35, 53

[24] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 9, 11, 19, 20, 21, 35, 54

[25] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 10, 21, 56

[26] Stephan R Sain. The nature of statistical learning theory, 1996. 10

[27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86 (11):2278–2324, 1998. 10

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 10, 11

[29] 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*, pages 770–778, 2016. 10, 12

[30] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 10

[31] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 11

[32] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 11

[33] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. A comprehensive study of deep video action recognition. *arXiv preprint arXiv:2012.06567*, 2020. 12

[34] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 12

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 12

[36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 12

[37] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 12

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, Miami, FL, USA, 2009. IEEE. 13, 35

[39] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 13

[40] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. *arXiv preprint arXiv:1712.04851*, 1(2):5, 2017. 13

[41] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 13

[42] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019. 13

[43] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, Seoul, Korea, 2019. IEEE. 13, 30, 36

[44] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, October 2021. 14, 56

[45] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 14, 44, 49, 55, 56

[46] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010. 16

[47] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press. 16, 28, 29

[48] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 17

[49] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 17

[50] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413, Long Beach, CA, USA, 2019. PMLR. 17, 26, 36, 45, 47, 49

[51] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 17, 18

[52] Han Zou, Yuxun Zhou, Jianfei Yang, Huihan Liu, Hari Prasanna Das, and Costas J Spanos. Consensus adversarial domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5997–6004, Honolulu, HI, USA, 2019. AAAI. 17, 18, 26

[53] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302, 2019. 17

[54] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105, Lille, France, 2015. PMLR. 18, 26, 36, 45, 47, 49

[55] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 18

[56] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Domain adaptation on the statistical manifold. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2481–2488, 2014. 18

[57] Waqas Sultani and Imran Saleemi. Human action recognition across datasets by foreground-weighted histogram decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 764–771, Columbus, Ohio, USA, 2014. IEEE. 19, 23

[58] Tiantian Xu, Fan Zhu, Edward K. Wong, and Yi Fang. Dual many-to-one-encoder-based transfer learning for cross-dataset human action recognition. *Image and Vision Computing*, 55:127–137, 2016. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2016.01.001. URL `https://www.sciencedirect.com/science/article/pii/S0262885616000020`. Handcrafted vs. Learned Representations for Human Action Recognition. 19

[59] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *BMVC*, volume 2, page 5, Newcastle, UK, 2018. BMVC. 19

[60] Xiao-Yu Zhang, Haichao Shi, Changsheng Li, Kai Zheng, Xiaobin Zhu, and Lixin Duan. Learning transferable self-attentive representations for action recognition in untrimmed videos with weak supervision. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9227–9234, Honolulu, HI, USA, 2019. AAAI. 19, 23

[61] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 19

[62] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Aligning correlation information for domain adaptation in action recognition, 2021. URL `https://arxiv.org/abs/2107.04932`. 19, 55, 56, 57, 58

[63] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 122–132, New York, NY, USA, 2020. IEEE. 19

[64] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9787–9795, Nashville, TN, USA, June 2021. IEEE. 19

[65] Jianming Lv, Kaijie Liu, and Shengfeng He. *Differentiated Learning for Multi-Modal Domain Adaptation*, page 1322–1330. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450386517. URL `https://doi.org/10.1145/3474085.3475660`. 19

[66] Fan Qi, Xiaoshan Yang, and Changsheng Xu. A unified framework for multimodal domain adaptation. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 429–437, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356657. doi: 10.1145/3240508.3240633. URL `https://doi.org/10.1145/3240508.3240633`. 19

[67] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 678–695. Springer, 2020. 19

[68] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 20, 54

[69] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 20, 53

[70] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *International Journal of Computer Vision*, 130(1):33–55, 2022. 21

[71] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, Glasgow, UK, 2018. Springer. 23, 24, 26, 28, 36, 38

[72] Yuecong Xu, Jianfei Yang, Haozhi Cao, Zhenghua Chen, Qi Li, and Kezhi Mao. Partial video domain adaptation with partial adversarial temporal attentive network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9332–9341, New York, NY, USA, 2021. IEEE. 23, 26, 28, 29, 31, 34, 35, 36, 37

[73] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8156–8164, Salt Lake City, Utah, USA, 2018. IEEE. 26, 28

[74] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2724–2732, Salt Lake City, Utah, USA, 2018. IEEE. 26

[75] Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2985–2994, Long Beach, CA, USA, 2019. IEEE. 26, 36

[76] Taotao Jing, Haifeng Xia, and Zhengming Ding. *Adaptively-Accumulated Knowledge Transfer for Partial Domain Adaptation*, page 1606–1614. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450379885. URL `https://doi.org/10.1145/3394171.3413986`. 24, 26, 28

[77] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039, Vienna, Austria, 2020. PMLR. 25, 27, 28, 32, 33, 34, 45, 49, 50, 55, 56, 58, 62

[78] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, Munich, Germany, 2018. Springer.

[79] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International conference on machine learning*, pages 5423–5432, Stockholm, Sweden, 2018. PMLR, PMLR. 25, 27, 28, 33

[80] Shuang Li, Chi Harold Liu, Binhui Xie, Limin Su, Zhengming Ding, and Gao Huang. Joint adversarial domain adaptation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 729–737, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3351070. URL `https://doi.org/10.1145/3343031.3351070`. 26

[81] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, Salt Lake City, Utah, USA, 2018. IEEE. 26, 36, 45, 47, 49, 55, 56, 57, 58

[82] Jianfei Yang, Han Zou, Shuxin Cao, Zhenghua Chen, and Lihua Xie. Mobileda: Toward edge-domain adaptation. *IEEE Internet of Things Journal*, 7(8):6909–6918, 2020.

[83] Jianfei Yang, Han Zou, Yuxun Zhou, Zhaoyang Zeng, and Lihua Xie. Mind the discriminability: Asymmetric adversarial domain adaptation. In *European Conference on Computer Vision*, pages 589–606. Springer, Springer, 2020.

[84] Jianfei Yang, Jiangang Yang, Shizheng Wang, Shuxin Cao, Han Zou, and Lihua Xie. Advancing imbalanced domain adaptation: Cluster-level discrepancy minimization with a comprehensive benchmark. *IEEE Transactions on Cybernetics*, pages 1–12, 2021. doi: 10.1109/TCYB.2021.3093888. 26

[85] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, Rui Zhao, and Zhenghua Chen. Multi-source video domain adaptation with temporal attentive moment alignment, 2021. URL `https://arxiv.org/abs/2109.09964`. 26

[86] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9944–9953, Seoul, Korea, 2019. IEEE. 27, 28, 33

[87] Rui Shu, Hung Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-t approach to unsupervised domain adaptation. In *International Conference on Learning Representations*, Vancouver, Canada, 2018. OpenReview.Net. URL `https://openreview.net/forum?id=H1q-TM-AW`. 27, 28, 33

[88] Baochen Sun and Kate Saenko. From virtual to reality: Fast adaptation of virtual object detectors to real domains. In *BMVC*, volume 1, page 3, Nottingham, United Kingdom, 2014. BMVA press. 27

[89] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223, Berlin, Heidelberg, 2007. Springer. 30

[90] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 33

[91] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006. 33

[92] Charles Elkan. Using the triangle inequality to accelerate k-means. In *Proceedings of the 20th international conference on Machine Learning (ICML-03)*, pages 147–153, Washington D.C., USA, 2003. PMLR. 33

[93] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. In *International Workshop on Deep Learning for Human Activity Recognition*, pages 70–84, Singapore, 2021. Springer. 35, 53

[94] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. *arXiv preprint arXiv:1712.04851*, 1(2):5, 2017. 35

[95] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:1–12, 2019. 35, 55

[96] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9772–9781, New York, NY, USA, 2021. IEEE. 35

[97] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625, Florence, Italy, 2012. Springer. 35

[98] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020. 44

[99] Muhammad Attique Khan, Kashif Javed, Sajid Ali Khan, Tanzila Saba, Usman Habib, Junaid Ali Khan, and Aaqif Afzaal Abbasi. Human action recognition using fusion of multiview and deep features: an application to video surveillance. *Multimedia tools and applications*, pages 1–27, 2020. 44

[100] Yu Kong, Zhiqiang Tao, and Yun Fu. Deep sequential context networks for action prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1481, 2017.

[101] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015.

[102] Michael S Ryoo, Thomas J Fuchs, Lu Xia, Jake K Aggarwal, and Larry Matthies. Robot-centric activity prediction from first-person videos: What will they do to me? In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pages 295–302, 2015.

[103] Jianfei Yang, Yuecong Xu, Haozhi Cao, Han Zou, and Lihua Xie. Deep learning and transfer learning for device-free human activity recognition: A survey. *Journal of Automation and Intelligence*, 1(1):100007, 2022. 44

[104] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 678–695. Springer, 2020. 44, 45, 47

[105] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, Rui Zhao, and Zhenghua Chen. Multi-source video domain adaptation with temporal attentive moment alignment. *arXiv preprint arXiv:2109.09964*, 2021.

[106] Yuecong Xu, Jianfei Yang, Haozhi Cao, Zhenghua Chen, Qi Li, and Kezhi Mao. Partial video domain adaptation with partial adversarial temporal attentive network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9332–9341, 2021. 44, 45, 47

[107] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Casting a bait for offline and online source-free domain adaptation. *arXiv preprint arXiv:2010.12427*, 2020. 45, 50

[108] Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. Source-free domain adaptation via avatar prototype generation and adaptation. *arXiv preprint arXiv:2106.15326*, 2021. 45, 50

[109] Riccardo Volpi, Diane Larlus, and Grégory Rogez. Continual adaptation of visual representations via domain randomization and meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4443–4453, 2021. 45, 47

[110] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, and Zhenghua Chen. Source-free video domain adaptation by learning temporal consistency for action recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 147–164. Springer, 2022. 45, 47

[111] Jogendra Nath Kundu, Rahul Mysore Venkatesh, Naveen Venkat, Ambareesh Revanur, and R Venkatesh Babu. Class-incremental domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 53–69. Springer, 2020. 45, 47, 49

[112] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 45, 47, 51, 52

[113] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526, 2017.

[114] Jaeyoo Park, Minsoo Kang, and Bohyung Han. Class-incremental learning for action recognition in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13698–13707, 2021. 45

[115] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 47

[116] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021. 47

[117] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021. 47

[118] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 47

[119] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetting learning in deep neural networks. *arXiv preprint arXiv:1607.00122*, 2016. 47, 51

[120] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018. 47

[121] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 47

[122] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1131–1140, 2020. 47

[123] Judy Hoffman, Trevor Darrell, and Kate Saenko. Continuous manifold based adaptation for evolving visual domains. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 867–874, 2014. 47

[124] Markus Wulfmeier, Alex Bewley, and Ingmar Posner. Incremental adversarial domain adaptation for continually changing environments. In *2018 IEEE International conference on robotics and automation (ICRA)*, pages 4489–4495. IEEE, 2018. 47

[125] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 47, 55, 56, 58

[126] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018. 48

[127] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013.

[128] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 48

[129] Sebastian Otálora, Manfredo Atzori, Vincent Andrearczyk, Amjad Khan, and Henning Müller. Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. *Frontiers in bioengineering and biotechnology*, 7:198, 2019. 48

[130] Riccardo Volpi and Vittorio Murino. Addressing model vulnerability to distributional shifts over image transformation sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7980–7989, 2019. 48

[131] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019. 48

[132] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. *arXiv preprint arXiv:2007.13003*, 2020. 48

[133] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*, 2021. 48

[134] Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. Dine: Domain adaptation from single and multiple black-box predictors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8003–8013, 2022. 49

[135] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 49, 50, 55, 56, 58

[136] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 51, 53

[137] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, Zhengguo Li, and Zhenghua Chen. Multi-source video domain adaptation with temporal attentive moment alignment network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 53, 54

[138] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 53, 54

[139] Abhay Yadav, Sohil Shah, Zheng Xu, David Jacobs, and Tom Goldstein. Stabilizing adversarial nets with prediction methods. *arXiv preprint arXiv:1705.07364*, 2017. 55

[140] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 56

[141] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 56

[142] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 56

[143] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013. 56