IMAGE AND VIDEO GENERATION VIA DEEP LEARNING

LIMING JIANG

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

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.

30/06/2023

Date

Liming Jiang

Liming Jiang

Supervisor Declaration Statement

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

30/06/2023

Date

Assoc Prof Chen Change Loy

Authorship Attribution Statement

This thesis contains material from four papers published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as the first author.

Chapter 3 is published as Liming Jiang, Ren Li, Wayne Wu, Chen Qian, Chen Change Loy. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

The contributions of the co-authors are as follows:

- I proposed the methodology, collected the data, designed and conducted the experiments, and prepared the manuscript.
- Ren Li assisted in conducting a part of the experiments and collecting the data.
- Wayne Wu provided useful suggestions to come up with the idea and revised the full draft.
- Chen Qian helped revise the manuscript and gave advice.
- Prof. Chen Change Loy figured out the initial research direction, offered insightful comments, and revised the manuscript carefully.

Chapter 4 is published as Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, Chen Change Loy. TSIT: A Simple and Versatile Framework for Image-to-Image Translation. In European Conference on Computer Vision (ECCV), 2020 (Spotlight).

The contributions of the co-authors are as follows:

- I proposed the methodology, collected the data, designed and conducted the experiments, and prepared the manuscript.
- Changxu Zhang assisted in conducting a part of the experiments.
- Mingyang Huang helped me come up with the methodology and revised the initial draft.
- Chunxiao Liu and Jianping Shi figured out the initial research direction and provided insightful comments.
- Prof. Chen Change Loy actively discussed with me about ideas and revised the manuscript carefully.

Chapter 5 is published as Liming Jiang, Bo Dai, Wayne Wu, Chen Change Loy. Focal Frequency Loss for Image Reconstruction and Synthesis. In IEEE International Conference on Computer Vision (ICCV), 2021.

The contributions of the co-authors are as follows:

- I figured out the research direction, proposed the methodology, collected the data, designed and conducted the experiments, and prepared the manuscript.
- Bo Dai and Wayne Wu actively discussed with me about methodology and experiments with useful comments. They also revised the initial draft.
- Prof. Chen Change Loy provided insightful suggestions and revised the manuscript carefully.

Chapter 6 is published as Liming Jiang, Bo Dai, Wayne Wu, Chen Change Loy. Deceive D: Adaptive Pseudo Augmentation for GAN Training with Limited Data. In Conference on Neural Information Processing Systems (NeurIPS), 2021.

The contributions of the co-authors are as follows:

- I figured out the research direction, proposed the methodology, collected the data, designed and conducted the experiments, and prepared the manuscript.
- Bo Dai and Wayne Wu actively discussed with me about methodology and experiments with useful comments. They also revised the initial draft.
- Prof. Chen Change Loy provided insightful suggestions and revised the manuscript carefully.

30/06/2023

Liming Jiang

Date

Liming Jiang

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Chen Change Loy, for his tremendous support, encouragement, and suggestions throughout my Ph.D. life. Without him, I could not have undertaken such a wonderful Ph.D. journey, which has been a turning point in my career and life. Prof. Loy always maintains a gentle, kind and patient attitude with his best care and sincerity. He not only provides me with a comfortable environment but also enough freedom to conduct research. He always treats me with courtesy, respects my decision, and never puts too much pressure on me. Anytime I discuss with him whether research, career, or daily life, his advice is invaluable. Prof. Loy helps me learn and sets a great example of what a true researcher should be. Words cannot express enough my gratitude to such a nice supervisor. A day as a teacher, a lifetime as a father.

I also wish to express my great appreciation to my close collaborators, Shuai Yang, Wayne Wu, and Bo Dai. Shuai Yang provides a lot of insightful suggestions for my research and also leads me to conduct a series of interesting studies these years. He is also my good friend in daily life. Wayne Wu is my mentor during my first year, who opens the door for me to enter my research field. I would not have gotten off to a good start without him. Bo Dai guides me to complete some interesting work and provides useful advice. He helps me come up with good ideas and revise my papers carefully. I also hope to thank my other collaborators, labmates, and peer Ph.D. students for their support in research discussion, data collection, and manuscript proofreading.

Besides, I am very thankful to my nice internship mentors, Chen Kong, Justin Theiss, Aayush Prakash, and Richard Newcombe at Meta; Krishna Kumar Singh, Richard Zhang, Yijun Li, and Jingwan (Cynthia) Lu at Adobe Research; Mingyang Huang, Chunxiao Liu, and Jianping Shi at SenseTime Research. They provide continuous support and useful suggestions during my wonderful internships, opening another door for my career.

In addition, I hope to thank my Thesis Advisory Committee (TAC) members, as well as all the anonymous reviewers for my papers and thesis. I also wish to thank the administrators in my school and the committee to support my scholarships to help me complete a smooth Ph.D. study.

Last but not least, I would like to express my profound gratitude to my family. They provide me with unceasing encouragement, support, and patience, as well as the warmest love. Without their accompany, the accomplishment of my Ph.D. would not have been possible.

"Try not to become a man of success, but rather try to become a man of value."

—Einstein, Albert

To my dear family

Contents

Abstract

Image and video generation aims at synthesizing high-fidelity visual data from random noise or based on certain conditions. Recent advances, especially Generative Adversarial Networks (GANs), have made remarkable success in various image and video generation tasks, showing the strong ability of deep neural networks to capture high dimensional distributions of visual data. Such progress in this field significantly pushes forward the development of Generative Artificial Intelligence (AI), receiving wide attention from the general public. Despite the immense success in image and video synthesis, several problems still exist to be carefully explored. This thesis aims to figure out the remaining challenges in this field and present efforts to address them through advanced deep learning techniques. This formulates four main studies to be covered.

Data is the essence of deep learning. A high-quality dataset is highly desirable for image and video generation, as well as its downstream applications. Besides, researchers usually pay much attention to improving generation quality but ignore the countermeasures to safeguard against the concerns raised by generated data like "Deepfakes". Different from others, the first attempt in this thesis is to construct a useful facial video dataset to facilitate the following research and prevent the negative impact of generated data by devising a better video manipulation method. DeeperForensics-1.0, a large-scale video dataset for real-world face forgery detection, is introduced. The ongoing effort is presented to counter "Deepfakes", which has sparked legitimate concerns, particularly on its potential for being misused and abused. It represents one of the most extensive datasets of the same kind, with $60,000$ videos constituted by a total of 17.6 million frames. Extensive real-world perturbations are applied to obtain a more challenging benchmark of larger scale and higher diversity. All source videos in DeeperForensics-1.0 are carefully collected, and fake videos are generated by a newly proposed end-to-end face swapping framework. The quality of generated videos outperforms those in existing datasets, validated by user studies. The benchmark features a hidden test set,

which contains manipulated videos achieving high deceptive scores. A comprehensive study is conducted that evaluates five representative detection baselines and makes a thorough analysis of different settings. This work verifies that designing a better video manipulation method can assist in face forensics.

After securing the potential countermeasures, the interest then shifts to proposing a unified framework for various generation tasks with a negligible sacrifice of quality, which has high practical value for real-world applications. Achieving this goal is non-trivial given the different natures of different tasks. Thus, previous studies usually exploit tailored modules for a specific form of application. This thesis devises a Two-Stream Image-to-image Translation (TSIT) framework, which is succinct yet readily adaptable to various tasks. The thesis unearths the importance of normalization layers and carefully designs a two-stream generative model with newly proposed feature transformations in a coarse-to-fine fashion. This allows multi-scale semantic structure information and style representation to be effectively captured and fused by the network, permitting TSIT to scale to various tasks in both unsupervised and supervised settings. No additional constraints (*e.g.*, cycle consistency) are needed, contributing to a very clean and simple method. Multimodal image synthesis with arbitrary style control is made possible. A systematic study compares TSIT with state-of-the-art task-specific baselines, verifying its effectiveness in both perceptual quality and quantitative evaluations.

Apart from the progress in the practical perspective of image and video generation, the thesis further wishes to tackle the remaining issues through a more fundamental and theoretical study. The third work in this thesis is the focal frequency loss (FFL), a novel frequency-level loss function that directly optimizes generative models in the frequency domain. The loss is complementary to existing spatial losses of diverse baselines varying in categories, network structures, and tasks. Despite the remarkable success of image reconstruction and synthesis thanks to the development of generative models, gaps could still exist between the real and generated images, especially in the frequency domain. The thesis shows that narrowing gaps in the frequency domain can ameliorate image reconstruction and synthesis quality further. The proposed FFL allows a model to adaptively focus on the frequency components that are hard to synthesize by down-weighting the easy ones. This objective function offers great impedance against the loss of important frequency information due to the inherent bias of neural networks. The thesis demonstrates

the versatility and effectiveness of FFL to improve popular models, such as VAE, pix2pix, and SPADE, in both perceptual quality and quantitative performance. Its potential on StyleGAN2 is further shown.

Attempts have been made to enhance the fidelity and diversity of synthesized data through both practical and theoretical aspects. However, current generative models like GANs typically require a large amount of training data to fully unleash their power, whereas collecting sufficient data samples is sometimes infeasible. Training generative models with fewer data while preserving synthesis quality remains underexplored. The thesis further introduces Adaptive Pseudo Augmentation (APA), a simple yet effective strategy for GAN training with limited data. Recent studies have shown that training GANs with limited data remains formidable due to discriminator overfitting, the underlying cause that impedes the generator's convergence. The introduced APA encourages healthy competition between the generator and the discriminator. As an alternative method to existing approaches that rely on standard data augmentations or model regularization, APA alleviates overfitting by employing the generator itself to augment the real data distribution with generated images, which deceives the discriminator adaptively. Extensive experiments showcase the effectiveness of APA in the low-data regime. A theoretical analysis is provided to examine the convergence and rationality of this new training strategy. APA is simple and effective. It can be added seamlessly to powerful contemporary GANs, such as StyleGAN2, with negligible computational cost.

Last but not least, the thesis discusses other relevant topics and envisions potential future work of image and video generation, *e.g.*, more advanced topics in video generation, the existing and future efforts on the new powerful diffusion models (DM), offering more insights into this research field.

List of Figures

List of Tables

Chapter 1

Introduction

In computer vision, image and video generation aims at synthesizing high-fidelity visual data from a random noise [4, 10, 11] or based on certain conditions, such as a class label [12, 13] or other data [3, 14, 15]. The former is called unconditional generation, while the latter is named conditional generation. Through advanced deep learning technology, image and video generation has grasped widespread attention over the past few years, and its practical applications have been widely expanded, *e.g.*, movie editing [16–18], style transfer [19, 20], image-to-image translation [3, 14, 21, 22], and face forensics [23, 24].

Recent advances [2, 10, 12, 25, 26], especially Generative Adversarial Networks (GANs) [10], have made remarkable success in various image and video generation tasks, showing the strong ability of deep neural networks to capture high dimensional distributions of visual data. Such progress in this field significantly pushes forward the development of Generative Artificial Intelligence (AI), receiving wide attention from the general public. Despite their immense success in image and video synthesis, several promising topics are still worth exploring. This thesis aims to figure out the remaining challenges in this field and present efforts in addressing them through deep learning. This chapter will cover the motivation and objectives, research scope, background and preliminaries, thesis contributions, and outline of the thesis.

1.1 Motivation and Objectives

The general research goal of this thesis is to provide attempts to address various unresolved open problems in the field of image and video generation. This section discusses the specific problems the thesis is going to tackle, as well as the motivation and objectives.

Data is the essence of deep learning. The amount of useful data usually matters for a higher upper-bound of model performance. A high-quality dataset is highly desirable for image and video generation, as well as its downstream applications. Besides, researchers usually pay much attention to improving generation quality but ignore the countermeasures to safeguard against the concerns raised by generated data. Different from others, the first attempt in this thesis aims to construct a useful video dataset to facilitate the following research and prevent the negative impact of generated data by devising a better video manipulation method. The thesis starts with human face data as face manipulation has grasped more and more attention. Recent advances [16–18] in automatic face swapping demonstrated impressive results. The cumbersome and tedious hand-crafted face editing efforts have been eschewed through advanced deep learning technology. Accordingly, the development of various face editing applications is expedited, contributing to both academia and industry. However, such strong face swapping techniques have also brought several moral and legitimate concerns. The abuse and misuse of these off-the-shelf softwares could lead to the spread of tampered videos and news, thus causing subsequent more serious harm. The conceivable perilous implications of current "Deepfakes" applications on the Internet have further set off alarm bells among the authorities and general public. Therefore, effective face forensics methods become a dire need to safeguard against these photorealistic "Deepfakes", especially in real-world scenarios where the video sources are rather unpredictable.

After securing the potential countermeasures against negative impact, the attention of the thesis is then shifted to another important and useful problem in image generation, *i.e.*, image-to-image translation, which aims at translating one image representation to another. The interest lies in proposing a generic and unified framework for various image-to-image translation tasks with a negligible sacrifice of quality. This framework is useful as it circumvents many cumbersome modifications when applied to different tasks, which is meaningful for many practical

applications. Nevertheless, achieving this goal is *non-trivial* given the different natures of different tasks. For example, paired data are usually unavailable in tasks like arbitrary style transfer. Under its unsupervised setting, image translation demands additional constraints, *e.g.*, cycle consistency [14, 19, 27, 28], semantic features [29], pixel gradients [30], pixel values [31]. On another note, supervised training pairs are available in semantic image synthesis (*i.e.*, translation from segmentation masks to images). This task is more data-dependent since the ground truth is usually required by typical reconstruction losses. Besides, specialized modules [32–36] are applied to maintain spatial coherence and resolution. Due to the different needs, previous approaches exploit some tailored modules for a specific form of application. It is difficult to cross-use these components or integrate them into a unified framework.

Apart from the progress in the practical perspective of image and video generation, this thesis further wishes to tackle the remaining issues through a more fundamental and theoretical study. The third work of the thesis aims at ameliorating the quality of image generation further through the frequency domain. The thesis aims to devise a novel frequency-level loss function that directly optimizes generative models in the frequency domain. The loss is complementary to existing spatial losses of diverse baselines varying in categories, network structures, and tasks. Despite the remarkable success of image reconstruction and synthesis, gaps could still exist between the real and generated images. It is observed that the gaps in the frequency domain between the real and fake images may be a common issue for various generative methods, albeit in slightly different forms. The observed gaps in the frequency domain may be imputed to some inherent bias of neural networks, *e.g.*, *spectral bias* [37–39], a learning bias of neural networks towards low-frequency functions. Thus, generative models tend to eschew frequency components that are hard to synthesize, *i.e.*, hard frequencies, and converge to an inferior point. Existing methods [2, 3, 34] usually adopt pixel losses in the spatial domain, while spatial domain losses hardly help a network find hard frequencies and synthesize them, in that every pixel shares the same significance for a certain frequency. In contrast, the thesis aims to carefully study the frequency domain gap and explore ways to ameliorate reconstruction and synthesis quality by narrowing this gap.

Thanks to the exploration of these useful applications and fundamental studies in the field of image and video generation, the performance of various generative

models can be evidently improved. However, current generative models like GANs typically require a large amount of training data to fully unleash their power. Training GANs with insufficient data tends to generate poor-quality images. In practice, collecting sufficient data samples for these GANs is sometimes infeasible, especially in domains where data are sparse and privacy-sensitive. To ease the practical deployment of powerful GANs, it is necessary to devise new strategies for training GANs with limited data while preserving the quality of synthesis. Recent studies have shown that the overfitting of the discriminator is the critical reason that impedes effective GAN training on limited data [40–43], rendering severe instability of training dynamics. Specifically, when the discriminator starts to overfit, the distributions of its outputs for real and generated samples gradually diverge from each other [41, 43], and its feedback to the generator becomes less informative. Consequently, the generator converges to an inferior point, compromising the quality of synthesized images. Recent solutions to this problem include the use of standard data augmentations, either conventional or differentiable, to real and generated images [41, 42, 44, 45] or applying an additional model regularization term [43]. Addressing the discriminator overfitting is still an open problem. Finding an alternative way to the aforementioned approaches can be interesting.

1.2 Research Scope

This thesis covers research on both image generation and video generation via advanced deep learning technology. Figure 1.1 illustrates the connections among the four primary studies (Chapters 3, 4, 5, 6). The thesis starts by discussing a useful dataset and two specific applications. First, a large-scale facial video dataset is built to explore an emerging application of video generation and manipulation, *i.e.*, face forensics. This work exploits high-quality generated videos to help detect falsified media in real-world scenarios. It facilitates the following research and prevents the potential negative impact. Besides, thanks to the state-of-the-art innovations in generative image modeling [4, 13, 34], another important application in image generation, *i.e.*, image-to-image translation, is carefully studied. The goal is to extend its potential and practical value further by designing a unified and versatile model. The thesis then provides insights into a fundamental and theoretical study of image generation, delving into optimization in the frequency domain for

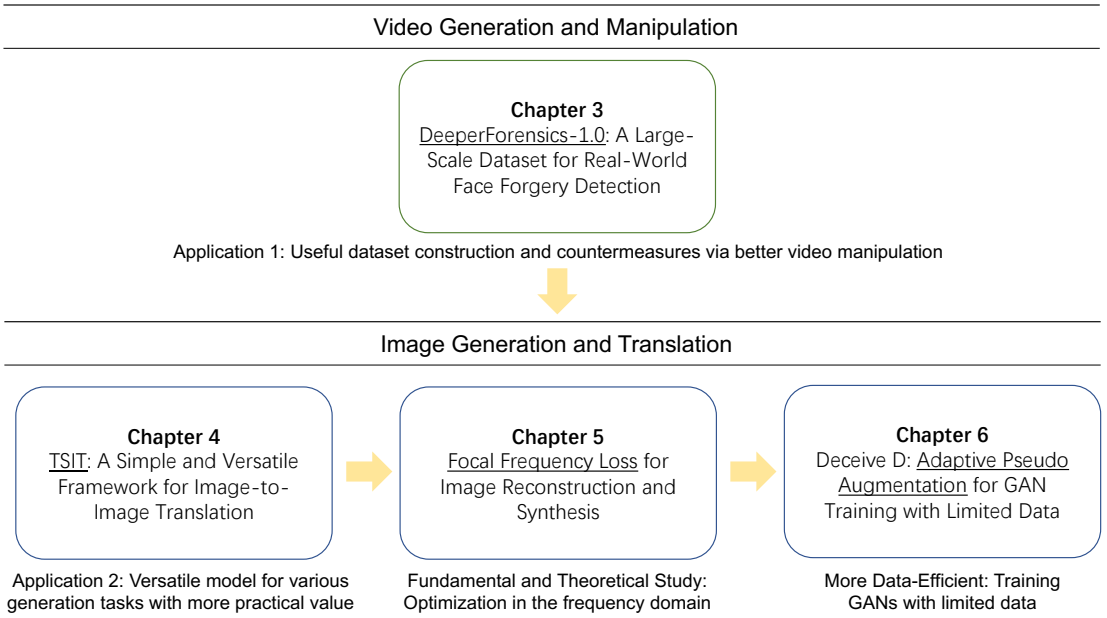

FIGURE 1.1: A diagram illustrating the connections among the four main studies (Chapters 3, 4, 5, 6) covered in this thesis.

the reconstruction and synthesis quality improvement, which is relatively less explored by academia. In addition, the thesis improves the training of GANs with lower data requirements yet with comparable quality, which further facilitates the practical deployment of the research in this thesis. In the future, any other studies closely related to image and video generation may be involved.

1.3 Background and Preliminaries

This section briefly introduces the background and formulations of the basic concepts relevant to image and video generation. The concept of generative models is introduced in Section 1.3.1, followed by generative adversarial networks, the major model explored in this thesis (Section 1.3.2).

1.3.1 Generative Models

In statistical classification, the approaches can be classified into two categories, *i.e.*, generative models and discriminative models. Different from the discriminative models that distinguish between different kinds of data instances, generative models

generate new data according to an existing dataset. Formally, given a set of data samples Y (and their corresponding labels X), generative models typically perform the following two tasks:

- Unconditional generation that synthesizes data unconditionally from a dataset, *i.e.*, capturing the marginal probability distribution $p(Y)$. The input of unconditional generation is usually a random noise sampled from a Gaussian distribution.
- Conditional generation that synthesizes data conditionally from a dataset, *i.e.*, capturing the joint probability distribution $p(Y, X)$. Such conditions can be the class labels of data, other data, *etc.*

Recent advances [1, 2, 10, 26, 46–48] of generative models are built on deep neural networks, showing impressive capability in capturing high-level latent representations of images and synthesizing new data. Two popular categories of generative models are autoencoders (AE) [1, 2] and GANs [10]. The vanilla AE [1] reconstructs images, aiming at learning latent codes in an unsupervised manner, typically for dimensional reduction and feature learning. Autoencoders have been widely used to generate images since the development of variational autoencoders (VAE) [2, 25]. Their applications have been extended to various tasks, *e.g.*, face manipulation [16, 17, 24]. Another category of generative models is GANs [10–12], details of which will be provided in Section 1.3.2. GANs are extensively applied in face generation [4, 5, 49], image-to-image translation [3, 14, 21, 50], style transfer [19, 20], and semantic image synthesis [33–35].

1.3.2 Generative Adversarial Networks

Generative adversarial networks [10] aim at capturing the real data distribution to synthesize new data. Two networks are trained alternately via an adversarial process: a generator G learns to produce new samples, and a discriminator D (*i.e.*, a binary classifier) predicts the probability that a sample comes from the real data rather than from G. Following [4, 10], the basic form of GANs is described using unconditional image generation, *i.e.*, synthesizing random samples from a noise input in the latent space. The noise is usually sampled from a Gaussian distribution.

The goal of GANs is to learn an ideal generated distribution p_g from the real data distribution p_{data}. Let $p_z(z)$ be the prior on the input noise variable. The mapping from the latent space to the image space can be denoted as $G(z)$. For sample x, $D(x)$ represents the estimated probability of x coming from the real data. Here, both G and D should be differentiable functions that are defined by the network parameters. To quantify the adversarial process, G and D play a minimax two-player game with the value function $V(G, D)$:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] . \quad (1.1)$$

Let the virtual training criterion [10] for the generator G be $C(G)$. The global minimum of $C(G)$ is achieved if and only if $p_g = p_{\text{data}}$, and the minimum value is $-\log 4$, as proved by [10]. This indicates that GANs can perfectly model the real data distribution if given sufficient capacity and time. In practice, researchers usually use a non-saturated form for G and train it to maximize $\log D(G(z))$ instead of minimizing $\log (1 - D(G(z)))$ to ensure a healthy gradient at the early training stage.

GANs are known to suffer from training instability [10, 40, 51, 52]. Various approaches have been proposed to stabilize the training and improve the quality of synthesis by minimizing different f-divergences of the real and fake distributions [53]. The saturated form of vanilla GAN [10] is theoretically proven to minimize the JS divergence [54] between the two distributions. LSGAN [55] and EBGAN [56] correspond to the optimizations of χ^2-divergence [57] and the total variation [58], respectively. On another note, WGAN [58] is designed for minimizing the Wasserstein distance.

State-of-the-art methods, such as PGGAN [49], BigGAN [13], StyleGAN [5], and StyleGAN2 [4], employ large-scale training with contemporary techniques, achieving photorealistic results. These methods have been extended to various tasks, including face generation [4, 5, 49], image editing [18, 50, 59], semantic image synthesis [33–35], image-to-image translation [3, 6, 14, 21, 22, 60], style transfer [19, 20, 61], and GAN inversion [22, 62, 63].

1.4 Thesis Contributions

This thesis presents the continuous efforts of developing innovative datasets and methods for image and video generation via deep learning. The contributions of the studies in this thesis (Chapters 3, 4, 5, 6) can are summarized as follows.

DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection (Chatper 3) in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020:

- The thesis introduces DeeperForensics-1.0 [24], a new dataset that is larger in scale than the existing ones, of high quality and rich diversity. To improve its quality, the thesis introduces a carefully designed data collection and a novel framework, DF-VAE, which effectively mitigate obvious fabricated effects of existing manipulated videos.
- The results of existing representative forgery detection methods are benchmarked on the introduced dataset, offering insights into the current status and future strategy in face forgery detection.

TSIT: A Simple and Versatile Framework for Image-to-Image Translation (Chapter 4) in European Conference on Computer Vision (ECCV), 2020 (Spotlight):

- The thesis introduces TSIT [21], a simple and versatile framework, which is effective for various image-to-image translation tasks. Despite the succinct design, TSIT is readily adaptable to various tasks and achieves compelling results.
- The good performance of TSIT is achieved by 1) *multi-scale* feature normalization (FADE and FAdaIN) scheme that captures *coarse-to-fine* structure and style information, and 2) a *two-stream* network design that integrates *both* content and style effectively, reducing artifacts and making multi-modal image synthesis possible.
- In comparison to several state-of-the-art task-specific baselines [20, 32–35, 64, 65], the proposed method achieves comparable or even better results in both perceptual quality and quantitative evaluations.

Focal Frequency Loss for Image Reconstruction and Synthesis (Chapter 5) in IEEE International Conference on Computer Vision (ICCV), 2021:

- The thesis introduces a novel focal frequency loss (FFL) [60], which directly optimizes generative models in the frequency domain. The loss is complementary to existing spatial losses, offering great impedance against the loss of hard frequencies due to the inherent crux of neural networks.
- Systematic experiments are conducted to demonstrate the versatility and effectiveness of the proposed loss for many popular image reconstruction and synthesis methods [1–4, 34] to ameliorate synthesis quality, outperforming relevant approaches [66, 67].
- The exact form of the focal frequency loss is not crucial. The thesis provides some variants and practical considerations for the flexibility.

Deceive D: Adaptive Pseudo Augmentation for GAN Training with Limited Data (Chapter 6) in Conference on Neural Information Processing Systems (NeurIPS), 2021:

- The thesis introduces a novel Adaptive Pseudo Augmentation (APA) [68] method for training GANs with limited data. This approach deceives the discriminator adaptively and mitigates the problem of discriminator overfitting. The proposed APA can be readily added to existing GAN training with negligible computational cost.
- Extensive experiments are conducted to showcase the effectiveness of APA for state-of-the-art GAN training with limited data. The results are comparable or even better than other types of solutions [41, 43]. APA is also complementary to existing methods based on standard data augmentations for gaining a further performance boost.
- The thesis theoretically connects APA with minimizing the JS divergence [54] between the smoothed data distribution and generated distribution, proving its convergence and rationality.

1.5 Outline of the Thesis

Chapter 1 has given a brief overview of image and video generation, the motivation and objectives, research scope, background and preliminaries, thesis contributions, and thesis outline. The rest of the thesis will be organized as follows.

Chapter 2 provides comprehensive literature reviews of image and video generation via deep learning, in the aspects of face forensics, image-to-image translation, frequency domain analysis, and training GANs with limited data.

Chapter 3 details the contributions to real-world face forensics via better video manipulation. The DeeperForensics-1.0 dataset and benchmark, as well as the proposed video manipulation method, are covered.

Chapter 4 introduces the proposed Two-Stream Image-to-image Translation (TSIT) framework. The network architecture and important modules of TSIT are detailed. The image translation results are shown and analyzed carefully.

Chapter 5 provides details on the formulation of focal frequency loss. The effectiveness and versatility of FFL to complement existing spatial losses are verified through systematic experiments on various networks and tasks.

Chapter 6 presents the APA strategy for training GANs with limited data. The effectiveness of APA is carefully showcased for state-of-the-art GAN training in the low-data regime both theoretically and empirically.

Chapter 7 concludes this thesis, discusses other relevant topics, and envisions future work, offering more insights into image and video generation via deep learning.

Chapter 2

Literature Review

This chapter reviews relevant literature on image and video generation via deep learning. The literature review provides an overview of four topics that correspond to four main studies discussed in this thesis: face forensics (Section 2.1), image-to-image translation (Section 2.2), frequency domain analysis (Section 2.3), and training GANs with limited data (Section 2.4).

2.1 Face Forensics

DeepFakes generation methods. The popularization of DeepFakes videos are attributed to the rapid development of generative models. Existing state-of-the-art generative models are mainly built on deep neural networks [1, 2, 10, 26, 46], showing impressive capability in capturing high-level latent representations of visual data and synthesizing new images. Two popular categories of generative models for face manipulation are autoencoders [1, 2] and GANs [10].

The vanilla AE [1] reconstructs images, aiming at learning latent codes in an unsupervised manner, typically for dimensional reduction and feature learning. Autoencoders have been widely used to generate images since the development of VAE [2, 25]. Extensive well-known off-the-shelf face manipulation softwares are based on autoencoders, *e.g.*, DeepFakes [16], DeepFaceLab [17, 69]. These methods tend to learn the identity information for face manipulation through the reconstruction process. However, they usually fit the specific domain and cannot scale

to multiple identities. The manipulation method DF-VAE (Section 3.2.2) for the DeeperForensics-1.0 dataset is based on variational autoencoders. DF-VAE is an end-to-end many-to-many face swapping method, which considers style matching and temporal continuity for video manipulation.

Another category of generative models is GAN [10–12], where a generator tries to fool a discriminator by refining the synthesized images continuously until the discriminator fails to perceive them as fakes. GAN has been extensively applied in face generation [4, 5, 49], image-to-image translation [3, 14, 21, 50, 60], style transfer [19, 20], and semantic image synthesis [21, 33–35, 60]. For face manipulation, the open-source DeepFakes software, faceswap-GAN [18], is a typical GAN-based method. It exploits adversarial losses to the denoising autoencoder and applies attention mechanisms to improve the clarity of the swapped faces. ReenactGAN [70] introduced the notion of boundary latent space for robust many-to-one face reenactment. Some recent GAN-based innovations were designed in the more challenging face manipulation context, *e.g.*, subject agnostic [71], occlusion aware [72].

Face forgery detection datasets. Building a dataset for forgery detection requires a huge amount of effort on data collection and manipulation. Early forgery detection datasets comprise images captured under highly restrictive conditions, *e.g.*, MICC_F2000 [73], Wild Web dataset [74], Realistic Tampering dataset [75].

Owing to the urgency in video-based face forgery detection, some prominent groups have devoted their efforts to create face forensics video datasets (see Table 2.1). UADFV [76] contains 98 videos, *i.e.*, 49 real videos from YouTube and 49 fake ones generated by FakeAPP [77]. DeepFake-TIMIT [78] manually selects 16 similar looking pairs of people from VidTIMIT [79] database. For each of the 32 subjects, they generate about 10 videos using low-quality and high-quality versions of faceswap-GAN [18], resulting in a total of 620 fake videos. Celeb-DF [80] includes 408 YouTube videos, mostly of celebrities, from which 795 fake videos are synthesized. FaceForensics++ [23] is the first large-scale face forensic dataset that consists of $4,000$ fake videos manipulated by four methods (*i.e.*, DeepFakes [16], Face2Face [81], FaceSwap [82], NeuralTextures [83])), and $1,000$ real videos from YouTube. Afterwards, Google joins FaceForensics++ and contributes Deep Fake Detection [84] dataset with $3,431$ real and fake videos from 28 actors. Recently, Facebook invites 66 individuals and builds the DFDC preview dataset [85], which includes $5,214$ original and tampered videos with three types of augmentations.

TABLE 2.1: The relevant datasets compared to DeeperForensics-1.0. Our dataset is an order of magnitude larger in scale than existing datasets *w.r.t.* both real and fake parts. We build a professional indoor environment to better control the important attributes of the collected data. 100 paid actors give consents to the use and manipulation of their faces by signing a formal agreement. We employ seven types of perturbations at five intensity levels, leading to 35 perturbations in total. The video may be subjected to a mixture of more than one perturbation. In contrast to prior works, we also introduce a new end-to-end high-fidelity face swapping method.

Dataset	Total videos	Ratio (real : fake)	Controlled Capture	Consented Actors	Perturbations (total number)	Perturbations (mixture)	New Method
UADFV [76]	98	1 : 1	×	–	–	×	×
DeepFake-TIMIT [78]	620	only fake	×	–	–	×	×
Celeb-DF [80]	1203	1 : 1.95	×	–	–	×	×
FaceForensics++ [23]	5000	1 : 4	×	–	2	×	×
Deep Fake Detection [84] (joins FaceForensics++)	3431	1 : 8.5	×	28	–	×	×
DFDC Preview Dataset [85]	5214	1 : 3.6	×	66	3	×	×
DeeperForensics-1.0 (Ours)	**60000**	5 : 1	✔	**100**	**35**	✔	✔

In comparison, our work in Chapter 3 invites 100 paid actors and collect high-resolution (1920×1080) source data with various poses, expressions, and illuminations. 3DMM blendshapes [86] are taken as reference to supplement some extremely exaggerated expressions. We get consents from all the actors for using and manipulating their faces. In contrast to prior works, we also propose a new end-to-end face swapping method (*i.e.*, DF-VAE in Section 3.2.2) and systematically apply seven types of perturbations to the fake videos at five intensity levels. The mixture of distortions to a single video makes our dataset better imitate real-world scenarios. Ultimately, we construct DeeperForensics-1.0 dataset, which contains up to $60,000$ high-quality videos with a total of 17.6 million frames.

Face forgery detection benchmarks. A new prominent benchmark, FaceForensics Benchmark [23], for facial manipulation detection has been proposed recently. The benchmark includes six image-level face forgery detection baselines [87–92]. Although FaceForensics Benchmark adds distortions to the videos by converting them into different compression rates, a deeper exploration of more perturbation types and their mixture is missing. Celeb-DF [80] also provides a face forgery detection benchmark including seven methods [76, 87, 89, 93–96] trained and tested on different datasets. In aforementioned benchmarks, the test set usually shares a similar distribution with the training set. Such an assumption inherently introduces biases and renders these methods impractical for face forgery detection in real-world settings with much more diverse and unknown fake videos.

In our benchmark of DeeperForensics-1.0 (Chapter 3), we introduce a challenging hidden test set with manipulated videos that achieve high deceptive scores in user studies, to better simulate *real-world* distribution. Various perturbations are analyzed to make our benchmark more comprehensive. In addition, we mainly exploit *video-level* forgery detection baselines [97–101]. Temporal information – a significant cue for video forgery detection besides single-frame quality – has been considered. We will elaborate our benchmark in Section 3.3.

2.2 Image-to-Image Translation

Taxonomy of image-to-image translation. Existing image-to-image translation methods can be generally grouped into two categories: unsupervised and supervised. With only unpaired data, unsupervised image-to-image translation problem is inherently ill-posed. Additional constraints are needed on *e.g.*, cycle consistency [14, 19, 27, 28], semantic features [29], pixel gradients [30], or pixel values [31]. In contrast, supervised methods, such as `pix2pix` [3], are more data-dependent, requiring well-annotated paired training samples. Subsequent approaches [32–35, 65] extend the supervised problem for generating high-resolution images or keeping effective semantic meaning.

Limited by learning only one-to-one mapping between two domains, some of the GAN-based methods [14, 19, 27, 28] suffer from generating images with low diversity. Recent studies explore more deeply into both multi-domain translation [50, 102] and multi-modal translation [20, 61, 103], significantly increasing generation diversity. MUNIT [20] is a representative method that disentangles the domain-invariant content and the domain-specific style representation, enriching the synthesized images. Multi-mapping translation is defined in a very recent work, DMIT [64], which is designed to capture the multi-modal image nature in each domain.

Existing image-to-image translation methods lack the scalability to adapt to different tasks under diverse difficult settings. Different demands of unsupervised and supervised settings oblige previous methods to exploit customized modules. Cross-using these components will be suboptimal due to either degradation in quality or introduction of additional constraints. It is non-trivial to integrate them into a

single framework and improve robustness. In Chapter 4, we design a two-stream network with newly proposed feature transformations inspired by [34] and [104]. Our method, TSIT, is succinct yet able to link various tasks.

Arbitrary style transfer. Style transfer is closely relevant to image-to-image translation in the unsupervised setting. Style transfer aims at retaining the content structure of an image, while manipulating its style representation adopted from other images. Classical methods [66, 105–107] gradually improve this task from optimization-based to real-time, allowing multiple style transfer during inference. Huang *et al.* . introduce AdaIN [104], an effective normalization strategy for arbitrary style transfer. Several studies [108–114] improve stylization via wavelet transforms [108], graph cuts [109], or iterative error-correction [114]. Besides, most collection-guided [20] style transfer methods are GAN-based [14, 19, 20, 27, 61, 64], showing impressive results.

Previous works usually consider either content or style information. In contrast, our framework in Chapter 4 succeeds in seeking a balance between content and style, and adaptively fuses them well. The proposed method achieves user-controllable multi-modal style manipulation by only a *single* model. Compared to customized style transfer methods, our approach achieves better synthesis quality in many scenarios including natural images, real-world scenes, and artistic paintings.

Semantic image synthesis. We define semantic image synthesis as in [34], aiming at synthesizing a photorealistic image from a semantic segmentation mask. Semantic image synthesis is a special form of supervised image-to-image translation. The domain gap of this task is large. Therefore, keeping effective semantic information to enhance fidelity without losing diversity is challenging.

`Pix2pix` [3] firstly adopts the conditional GAN [12] in semantic image synthesis. Pix2pixHD [33] contains a multi-scale generator and multi-scale discriminators to generate high-resolution images. SPADE [34] takes a noise map as input, and resizes the semantic label map for modulating the activations in normalization layers by a learned affine transformation. CC-FPSE [35] employs a weight prediction network for generator. A semantics-embedding discriminator is used to enhance fine details and semantic alignments between the generated samples and the input semantic layouts. In addition to these GAN-based methods, CRN [32] applies a cascaded refinement network with regression loss as the supervision. SIMS [65] is

a semi-parametric method, retrieving fragments from a memory bank and refining the canvas by a refinement network.

Different from prior works, we design a symmetrical two-stream framework in Chapter 4. The network learns feature-level semantic structure information and style representation instead of directly resizing the input mask like SPADE [34]. Coarse-to-fine feature representations are learned by neural networks, adaptively keeping high fidelity without diminishing diversity.

2.3 Frequency Domain Analysis

Image reconstruction and synthesis. Autoencoders [1, 2] and GANs [10] are two popular models for image reconstruction and synthesis. The vanilla AE [1] aims at learning latent codes while reconstructing images. It is typically used for dimensionality reduction and feature learning. Autoencoders have been widely used to generate images since the development of variational autoencoders [2, 25]. Their applications have been extended to various tasks, *e.g.*, face manipulation [16, 17, 24]. GAN [10–12], on the other hand, is extensively applied in face generation [4, 5, 49], image-to-image translation [3, 14, 21, 50], style transfer [19, 20], and semantic image synthesis [33–35]. Existing approaches usually apply spatial domain loss functions, *e.g.*, perceptual loss [66], to improve quality while seldom consider optimization in the frequency domain. Spectral regularization [67] presents a preliminary attempt. Different from [66, 67], the proposed focal frequency loss in Chapter 5 dynamically focuses the model on hard frequencies by down-weighting the easy ones and ameliorates image quality through the frequency domain directly. Some concurrent works include [115–117].

Frequency domain analysis of neural networks. In addition to the studies [37–39, 118] discussed in Section 5.1, we highlight some recent works that analyze neural networks from the frequency domain aspect. Using coordinate-based MLPs, Fourier features [39, 119] and positional encoding [38, 120] are adopted to recover missing high frequencies in single image regression problems. Besides, several studies have incorporated frequency analysis with network compression [121–125] and feature reduction [126, 127] to accelerate the training and inference of networks. The application areas of the frequency domain analysis have been further

extended, including media forensics [128–131], super-resolution [132, 133], generalization analysis [134], magnetic resonance imaging [135], image rescaling [136], *etc.* Despite the wide exploration of various problems, improving reconstruction and synthesis quality via the frequency domain remains much less explored.

Hard example processing. Hard example processing is widely explored in object detection and image classification to address the class imbalance problem. A common solution is to use a bootstrapping technique called hard example mining [137, 138], where a representative method is online hard example mining (OHEM) [137]. The training examples are sampled following the current loss of each example to modify the stochastic gradient descent. The model is encouraged to learn hard examples more to boost performance. An alternative solution is focal loss [139], which is a scaled cross-entropy loss. The scaling factor down-weights the contribution of easy examples during training so that a model can focus on learning hard examples. The proposed focal frequency loss in Chapter 5 is inspired by these techniques.

2.4 Training GANs with Limited Data

Generative adversarial networks. GANs [10–12, 140] adopt an adversarial training scheme, where a generator keeps refining its capability in synthesizing images to compete with a discriminator (*i.e.*, a binary classifier) until the discriminator fails to classify the generated samples as fakes. GANs are known to suffer from training instability [10, 40, 51, 52]. Various approaches have been proposed to stabilize the training and improve the quality of synthesis by minimizing different f-divergences of the real and fake distributions [53]. The saturated form of vanilla GAN [10] is theoretically proven to minimize the JS divergence [54] between the two distributions. LSGAN [55] and EBGAN [56] correspond to the optimizations of χ^2-divergence [57] and the total variation [58], respectively. On another note, WGAN [58] is designed for minimizing the Wasserstein distance.

State-of-the-art methods, such as PGGAN [49], BigGAN [13], StyleGAN [5], and StyleGAN2 [4], employ large-scale training with contemporary techniques, achieving photorealistic results. These methods have been extended to various tasks, including face generation [4, 5, 49], image editing [18, 50, 59], semantic image

synthesis [33–35], image-to-image translation [3, 6, 14, 21, 22, 60], style transfer [19, 20, 61], and GAN inversion [22, 62, 63]. Despite the remarkable success, the performance of GANs relies heavily on the amount of training data.

Training GANs in the low-data regime. The significance and difficulty of training GANs with limited data have been attracting attention from many researchers recently. The issue of data insufficiency tends to cause overfitting in the discriminator [41, 42, 141], which in turn deteriorates the stability of training dynamics in GANs, compromising the quality of generated images.

Many recent studies [41, 42, 44, 45, 142, 143] propose to apply standard data augmentations for GAN training to enrich the diversity of the dataset to mitigate the overfitting of the discriminator. For instance, DiffAugment (DA) [42] adopts the same differentiable augmentation to both real and fake images for the generator and the discriminator without manipulating the target distribution. Adaptive discriminator augmentation (ADA) [41] shares a similar idea with DA, while it further devises an adaptive approach that controls the strength of data augmentations adaptively. In Chapter 6, we extend the study of such an adaptive approach.

Another type of solution is model regularization. Previous efforts on regularizing GANs include adding noise to the inputs of the discriminator [40, 144, 145], gradient penalty [52, 146, 147], one-sided label smoothing [51], spectral normalization [148], label noise [149], *etc.* These methods are designed for stabilizing training or preventing mode collapse [51]. The essence of their goals could be considered similar to our method since training GANs in the low-data regime exhibits similar behaviors as previously observed in early GANs with sufficient data. Under the limited data setting, a very recent study proposes an LC-regularization term [43] to regulate the discriminator predictions using two exponential moving average variables that track the discriminator outputs throughout training.

Our work in Chapter 6 explores an alternative solution from a different perspective, which is also complementary to previous approaches based on standard data augmentations.

Chapter 3

DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection

3.1 Introduction

Data is the essence of deep learning. The amount of useful data usually matters for a higher upper-bound of model performance. A high-quality dataset is highly desirable for image and video generation, as well as its downstream applications. Besides, researchers usually pay much attention to improving generation quality but ignore the countermeasures to safeguard against the concerns raised by generated data. Different from others, our first attempt in this chapter aims to construct a useful facial video dataset to facilitate the following research and prevent the negative impact of generated data by devising a better video manipulation method.

Face swapping has become an emerging topic in computer vision and graphics. Indeed, many works [16–18] on automatic face swapping have been proposed in recent years. These efforts have circumvented the cumbersome and tedious manual face editing processes, hence expediting the advancement in face editing. At the

* The work in this chapter has been published in [24].

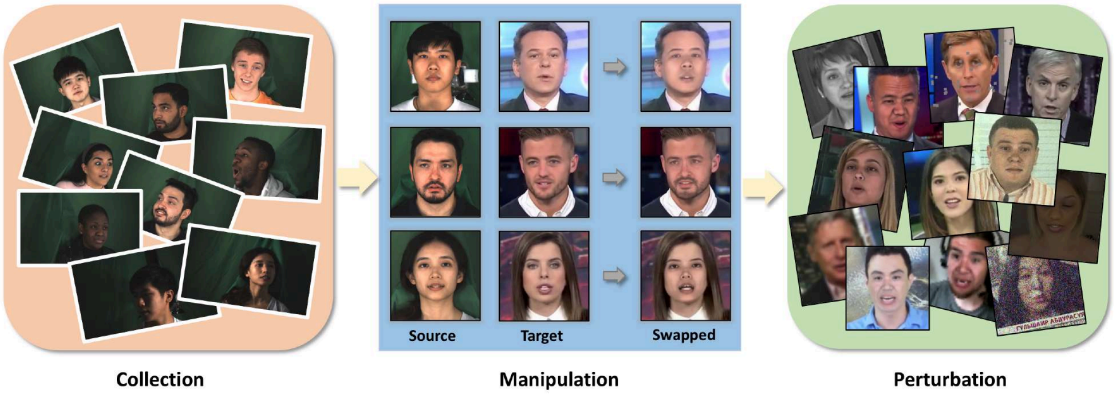

FIGURE 3.1: DeeperForensics-1.0 is a new large-scale dataset for *real-world* face forgery detection.

same time, such enabling technology has sparked legitimate concerns, particularly on its potential for being misused and abused. The popularization of "Deepfakes" on the internet has further set off alarm bells among the general public and authorities, in view of the conceivable perilous implications. Accordingly, there is a dire need for countermeasures to be in place promptly, particularly innovations that can effectively detect videos that have been manipulated.

Working towards forgery detection, various groups have contributed datasets (*e.g.*, FaceForensics++ [23], Deep Fake Detection [84] and DFDC [85]) comprising manipulated video footages. The availability of these datasets has undoubtedly provided essential avenues for research into forgery detection. Nonetheless, the aforementioned datasets suffer several drawbacks. Videos in these datasets are either of a small number, of low quality, or overly artificial. Understandably, these datasets are inadequate to train a good model for effective forgery detection in *real-world* scenarios. This is particularly true when current advances in human face editing are able to produce extremely realistic videos, rendering forgery detection a highly challenging task. On another note, we observe high similarity between training and test videos, in terms of their distribution, in certain works [23, 80]. Their actual efficacy in detecting *real-world* face forgery cases, which are much more variable and unpredictable, remains to be further elucidated.

We believe that forgery detection models can only be enhanced when trained with a dataset that is exhaustive enough to encompass as many potential real-world variations as possible. To this end, we propose a large-scale dataset named DeeperForensics-1.0 consisting of 60,000 videos with a total of 17.6 million frames

for real-world face forgery detection. The main steps of our dataset construction are shown in Figure 3.1. We set forth three yardsticks when constructing this dataset: 1) *Quality.* The dataset shall contain videos more realistic and much closer to the distribution of real-world detection scenarios. (Section 3.2.1 and 3.2.2) 2) *Scale.* The dataset shall be made up of a large-scale video sets. (Section 3.2.3) 3) *Diversity.* There shall be sufficient variations in the video footages (*e.g.*, compression, blurry, transmission errors) to match those that may be encountered in the real world (Section 3.2.3).

The primary challenge in the preparation of this dataset is the lack of good-quality video footages. Specifically, most publicly available videos are shot under an unconstrained environment resulting in large variations, including but not limited to suboptimal illumination, large occlusion of the target faces, and extreme head poses. Importantly, the lack of official informed consents from the video subjects precludes the use of these videos, even for non-commercial purposes. On the other hand, while some videos of manipulated faces are deceptively real, a larger number remains easily distinguishable by human eyes. The latter is often caused by model negligence towards appearance variations or temporal differences, leading to preposterous and incongruous results.

We approach the aforementioned challenge from two perspectives. 1) Collecting fresh face data from 100 individuals with informed consents (Section 3.2.1). 2) Devising a novel method, DeepFake Variational Auto-Encoder (DF-VAE), to enhance existing videos (Section 3.2.2). In addition, we introduce diversity into the video footages through deliberate addition of distortions and perturbations, simulating real-world scenarios. We collate the newly collected data and the DF-VAE-modified videos into the DeeperForensics-1.0 dataset, with the aim of further expanding it gradually over time. We benchmark five representative open-source forgery detection methods using our dataset as well as a hidden test set containing manipulated videos that achieve high deceptive ranking in user studies.

We summarize our contributions as follows: 1) We propose DeeperForensics-1.0, a new dataset that is larger in scale than the existing ones, of high quality and rich diversity. To improve its quality, we introduce a carefully designed data collection and a novel framework, DF-VAE, which effectively mitigate obvious fabricated effects of existing manipulated videos. The DeeperForensics-1.0 dataset shall facilitate future research in forgery detection of human faces in real-world scenarios.

FIGURE 3.2: Comparison of using only YouTube video and the collected video as source data, with the same method and setting.

2) We benchmark results of existing representative forgery detection methods on our dataset, offering insights into the current status and future strategy in face forgery detection.

3.2 A New Large-Scale Face Forensics Dataset

The main contribution of this work is a new large-scale dataset for real-world face forgery detection, DeeperForensics-1.0, which provides an alternative to existing databases. DeeperForensics-1.0 consists of 60,000 videos with 17.6 million frames in total, including 50,000 original collected videos and 10,000 manipulated videos. To construct a dataset more suitable for real-world face forgery detection, we design this dataset with careful consideration of *quality*, *scale*, and *diversity*. In Section 3.2.1 and 3.2.2, we will discuss the details of data collection and methodology (*i.e.*, DF-VAE) to improve *quality*. In Section 3.2.3, we will show how to ensure large *scale* and high *diversity* of DeeperForensics-1.0.

3.2.1 Data Collection

Source data is the first factor that highly affects *quality*. Taking results in Figure 3.2 as an example, the source data collection increases the robustness of our

FIGURE 3.3: Diversity in identities, poses, expressions, and illuminations in our collected source data.

face swapping method to extreme poses, since videos on the internet usually have limited head pose variations.

We refer to the identity in the driving video as the "target" face and the identity of the face that is swapped onto the driving video as the "source" face. Different from previous works, we find that the source faces play a much more critical role than the target faces in building a high-quality dataset. Specifically, the expressions, poses, and lighting conditions of source faces should be much richer in order to perform robust face swapping. Hence, our data collection mainly focuses on source face videos. Figure 3.3 shows the diversity in different attributes of our data collection.

We invite 100 paid actors to record the source videos. Similar to [84, 85], we obtain consents from all the actors for using and manipulating their faces to avoid the portrait right issues. The participants are carefully selected to ensure variability in genders, ages, skin colors, and nationalities. We maintain a roughly equal proportion *w.r.t.* each of the attributes above. In particular, we invite 55 males and 45 females from 26 countries. Their ages range from 20 to 45 years old to match the most common age group appearing on real-world videos. The actors have four typical skin tones: *white*, *black*, *yellow*, *brown*, with ratio 1:1:1:1. All faces are clean without glasses or decorations.

Different from previous data collection in the wild (see Table 2.1), we build a professional *indoor* environment for a more controllable data collection. We only use the facial regions (detected and cropped by LAB [150]) of the source data, so we can neglect the background. We set seven HD cameras from different angles: front, left, left-front, right, right-front, oblique-above, oblique-below. The resolution of our recorded videos is high (1920×1080). We train the actors in advance to keep the collection process smooth. We request the actors to turn their heads and speak

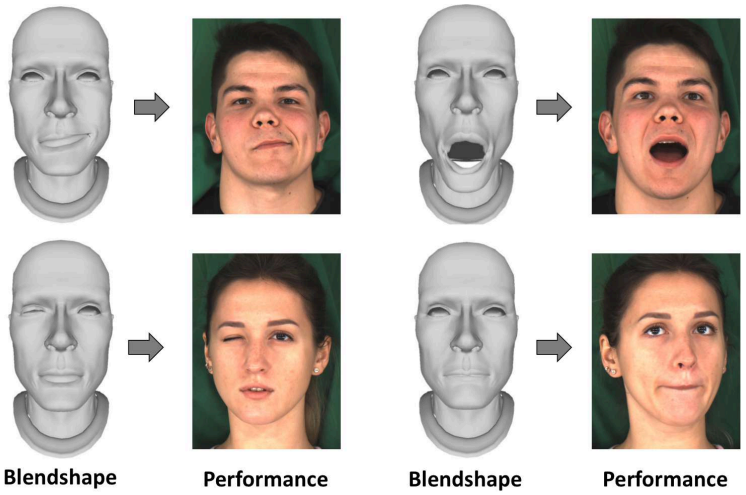

FIGURE 3.4: Examples of 3DMM blendshapes in our data collection.

naturally with eight expressions: neutral, angry, happy, sad, surprise, contempt, disgust, fear. The head poses range from $-90°$ to $+90°$. Furthermore, the actors are asked to perform 53 expressions defined in 3DMM blendshapes [86] (see Figure 3.4) to supplement some extremely exaggerated expressions. When performing 3DMM blendshapes, the actors also speak naturally to avoid excessive frames that show a closed mouth.

In addition to expressions and poses, we systematically set nine lighting conditions from various directions: uniform, left, top-left, bottom-left, right, top-right, bottom-right, top, bottom. The actors are only asked to turn their heads under uniform illumination, so the lighting remains unchanged on specific facial regions to avoid many duplicated data samples recorded by the cameras set at different angles. In the end, our collected data contain over $50,000$ videos with a total of 12.6 million frames – an order of magnitude more than existing datasets.

3.2.2 DeepFake Variational Auto-Encoder

To tackle low visual *quality* problems of previous works, we consider three key requirements in formulating a high-fidelity face swapping method: 1) It should be general and scalable for us to generate large number of videos with high quality. 2) The problem of face style mismatch caused by appearance variations need to be addressed. Some failure cases of existing methods are shown in Figure 3.5. 3) Temporal continuity of generated videos should be taken into consideration.

FIGURE 3.5: Examples of style mismatch problems in prominent face forensics datasets.

Based on the aforementioned requirements, we propose DeepFake Variational Auto-Encoder (DF-VAE), a novel learning-based face swapping framework. DF-VAE consists of three main parts, namely a structure extraction module, a disentangled module, and a fusion module. We will give a brief and intuitive understanding of the DF-VAE framework below.

Disentanglement of structure and appearance. The first step of our method is face reenactment – animating the source face with similar expression as the target face, without any paired data. Face swapping is considered as a subsequent step of face reenactment that performs fusion between the reenacted face and the target background. For robust and scalable face reenactment, we should cleanly disentangle structure (*i.e.*, expression and pose) and appearance representation (*i.e.*, texture, skin color, *etc.*) of a face. This disentanglement is rather difficult because structure and appearance representation are far from independent. We describe our solution as follows.

Let $\mathbf{x}_{1:T} \equiv \{x_1, x_2, ..., x_T\} \in X$ be a sequence of source face video frames, and $\mathbf{y}_{1:T} \equiv \{y_1, y_2, ..., y_T\} \in Y$ be the sequence of corresponding target face video frames. We first simplify our problem and only consider two specific snapshots at time t, x_t and y_t. Let $\tilde{x}_t$, $\tilde{y}_t$, d_t represent the reconstructed source face, the reconstructed target face, and the reenacted face, respectively.

Consider the reconstruction procedure of the source face x_t. Let s_x denotes the structure representation and a_x denotes the appearance information. The face generator can be depicted as the posteriori estimate $p_\theta(x_t|s_x, a_x)$. The solution of our reconstruction goal, marginal log-likelihood $\tilde{x}_t \sim \log p_\theta(x_t)$, by a common

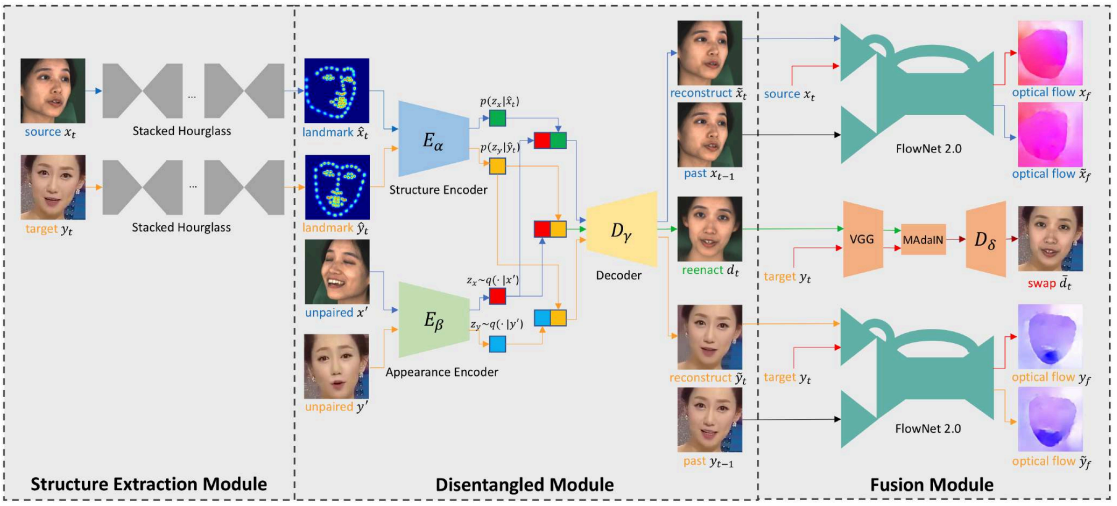

FIGURE 3.6: The main framework of DeepFake Variational Auto-Encoder. In training, we reconstruct the source and target faces in blue and orange arrows, respectively, by extracting landmarks and constructing an unpaired sample as the condition. Optical flow differences are minimized after reconstruction to improve temporal continuity. In inference, we swap the latent codes and get the reenacted face in green arrows. Subsequent MAdaIN module fuses the reenacted face and the original background resulting in the swapped face.

VAE [2] can be written as:

$$\begin{aligned} \log p_\theta(x_t) = D_{KL}\left(q_\phi\left(s_x, a_x|x_t\right) \| p_\theta\left(s_x, a_x|x_t\right)\right) \\ + L\left(\theta, \phi; x_t\right), \end{aligned} \tag{3.1}$$

where q_ϕ is an approximate posterior to achieve the evidence lower bound (ELBO) in the intractable case, and the second RHS term $L(\theta, \phi; x_t)$ is the variational lower bound *w.r.t.* both the variational parameters ϕ and generative parameters θ.

In Eq. (3.1), we assume that both s_x and a_x are latent priors computed by the same posterior x_t. However, the separation of these two variables in the latent space is rather difficult without additional conditions. Therefore, we employ a simple yet effective approach to disentangle these two variables.

The blue arrows in Figure 3.6 demonstrate the reconstruction procedure of the source face x_t. Instead of feeding a single source face x_t, we sample another source face x' to construct unpaired data in the source domain. To make the structure representation more evident, we use the stacked hourglass networks [151] to extract landmarks of x_t in the structure extraction module and get the heatmap $\hat{x}_t$. Then we feed the heatmap $\hat{x}_t$ to the Structure Encoder E_α, and x' to the Appearance

Encoder E_β. We concatenate the latent representations (small cubes in red and green) and feed it to the Decoder D_γ. Finally, we get the reconstructed face $\tilde{x}_t$, *i.e.*, marginal log-likelihood of x_t.

Therefore, the latent structure representation s_x in Eq. (3.1) becomes a more evident heatmap representation $\hat{x}_t$, which is introduced as a new condition. The unpaired sample x' with the same identity *w.r.t.* x_t is another condition, being a substitute for a_x. Eq. (3.1) can be rewritten as a conditional log-likelihood:

$$\begin{aligned}\log p_\theta\left(x_t|\hat{x}_t, x'\right) = D_{KL}\left(q_\phi\left(z_x|x_t, \hat{x}_t, x'\right)\|p_\theta\left(z_x|x_t, \hat{x}_t, x'\right)\right) \\ + L\left(\theta, \phi; x_t, \hat{x}_t, x'\right),\end{aligned} \tag{3.2}$$

The first RHS term KL-divergence is non-negative, we get:

$$\begin{aligned}&\log p_\theta\left(x_t|\hat{x}_t, x'\right) \geq L(\theta, \phi; x_t, \hat{x}_t, x') \\ &= \mathbb{E}_{q_\phi(z_x|x_t, \hat{x}_t, x')}\left[-\log q_\phi\left(z_x|x_t, \hat{x}_t, x'\right) + \log p_\theta\left(x_t, z_x|\hat{x}_t, x'\right)\right],\end{aligned} \tag{3.3}$$

and $L(\theta, \phi; x_t, \hat{x}_t, x')$ can also be written as:

$$\begin{aligned}L\left(\theta, \phi; x_t, \hat{x}_t, x'\right) = &- D_{KL}\left(q_\phi\left(z_x|x_t, \hat{x}_t, x'\right)\|p_\theta\left(z_x|\hat{x}_t, x'\right)\right) \\ &+ \mathbb{E}_{q_\phi(z_x|x_t, \hat{x}_t, x')}\left[\log p_\theta\left(x_t|z_x, \hat{x}_t, x'\right)\right].\end{aligned} \tag{3.4}$$

We let the variational approximate posterior be a multivariate Gaussian with a diagonal covariance structure:

$$\log q_\phi\left(z_x|x_t, \hat{x}_t, x'\right) \equiv \log \mathcal{N}\left(z_x; \mu, \sigma^2\mathbf{I}\right), \tag{3.5}$$

where $\mathbf{I}$ is an identity matrix. Exploiting the reparameterization trick [2], the non-differentiable operation of sampling can become differentiable by an auxiliary variable with independent marginal. In this case, $z_x \sim q_\phi\left(z_x|x_t, \hat{x}_t, x'\right)$ is implemented by $z_x = \mu + \sigma\epsilon$ where ϵ is an auxiliary noise variable $\epsilon \sim \mathcal{N}(0, 1)$. Finally, the approximate posterior $q_\phi(z_x|x_t, \hat{x}_t, x')$ is estimated by the separated encoders, Structure Encoder E_α and Appearance Encoder E_β, in an end-to-end training process by standard gradient descent.

We discuss the whole workflow of reconstructing the source face. In the target face domain, the reconstruction procedure is the same, as shown by orange arrows in Figure 3.6.

FIGURE 3.7: Comparison of the swapped face styles without or with MAdaIN module.

During training, the network learns structure and appearance information in both the source and the target domains. It is noteworthy that even if both y_t and x' belong to arbitrary identities, our effective disentangled module is capable of learning meaningful structure and appearance information of each identity. During inference, we concatenate the appearance prior of x' and the structure prior of y_t (small cubes in red and orange) in the latent space, and the reconstructed face d_t shares the same structure with y_t and keeps the appearance of x'. Our framework allows concatenations of structure and appearance latent codes extracted from arbitrary identities in inference and permits *many-to-many face reenactment*.

In summary, DF-VAE is a new conditional variational auto-encoder [25] with robustness and scalability. It conditions on two posteriors in different domains. In the disentangled module, the separated design of two encoders E_α and E_β, the explicit structure heatmap, and the unpaired data construction jointly force E_α to learn structure information and E_β to learn appearance information.

Style matching and fusion. To fix the obvious style mismatch problems shown in Figure 3.5, we introduce a masked adaptive instance normalization (MAdaIN) module. We place a typical AdaIN [104] network after the reenacted face d_t. In the face swapping scenario, we only need to adjust the style of the face area and use the original background. Therefore, we use a mask m_t to guide AdaIN [104] network to focus on style matching of the face area. To avoid boundary artifacts, we apply Gaussian Blur to m_t and get the blurred mask m_t^b.

In our face swapping context, d_t is the content input of MAdaIN, y_t is the style input. MAdaIN adaptively computes the affine parameters from the face area of

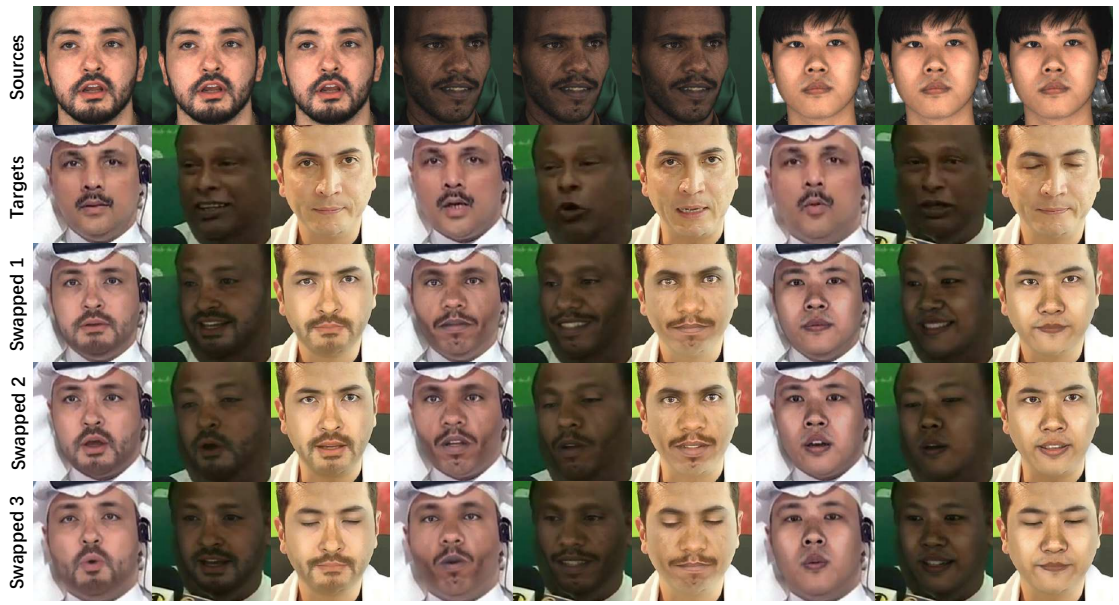

FIGURE 3.8: Many-to-many (three-to-three) face swapping by a **single** model with obvious reduction of style mismatch problems. This figure shows the results between three source identities and three target identities. The whole process is end-to-end.

the style input:

$$\mathrm{MAdaIN}\left(\mathrm{c},\mathrm{s}\right)=\sigma\left(\mathrm{s}\right)\left(\frac{\mathrm{c}-\mu\left(\mathrm{c}\right)}{\sigma\left(\mathrm{c}\right)}\right)+\mu\left(\mathrm{s}\right), \tag{3.6}$$

where $c = m_t^b \cdot d_t$, $s = m_t^b \cdot y_t$. With the very low-cost MAdaIN module, we reconstruct d_t again by Decoder D_δ. The blurred mask m_t^b is used again to fuse the reconstructed image with the background of y_t. At last, we get the swapped face $\overline{d}_t$. Figure 3.7 shows the effectiveness of MAdaIN module for style matching and fusion.

The MAdaIN module is jointly trained with the disentangled module in an end-to-end manner. Thus, by a *single* model, DF-VAE can perform *many-to-many face swapping* with obvious reduction of style mismatch and facial boundary artifacts (see Figure 3.8 for the face swapping between three source identities and three target identities). Even if there are multiple identities in both the source domain and the target domain, the quality of face swapping does not degrade.

Temporal consistency constraint. Temporal discontinuity of fake videos leads to obvious flickering of the face area, making them very easy to be spotted by

forgery detection methods and human eyes. To improve temporal continuity, we let the disentangled module to learn temporal information of both the source face and the target face.

For simplification, we make a Markov assumption that the generation of the frame at time t sequentially depends on its previous P frames $\mathbf{x}_{(t-p):(t-1)}$. In our experiment, we set $P = 1$ to balance quality improvement and training time.

In order to build the relationship between a current frame and previous ones, we further make an intuitive assumption that the optical flows should remain unchanged after reconstruction. We use FlowNet 2.0 [152] to estimate the optical flow $\tilde{x}_f$ *w.r.t.* $\tilde{x}_t$ and x_{t-1}, x_f *w.r.t.* x_t and x_{t-1}. Since face swapping is sensitive to minor facial details which can be greatly affected by flow estimation, we do not warp x_{t-1} by the estimated flow like [15]. Instead, we minimize the difference between $\tilde{x}_f$ and x_f to improve temporal continuity while keeping stable facial detail generation. To this end, we propose a new temporal consistency constraint, which can be written as:

$$L_{temporal} = \frac{1}{CHW}\|\tilde{x}_f - x_f\|_1, \tag{3.7}$$

where $C = 2$ for a common form of optical flow.

We only discuss the temporal continuity *w.r.t.* the source face in this section because the case of the target face is the same. If multiple identities exist in one domain, temporal information of all these identities can be learned in an end-to-end manner.

3.2.3 Scale and Diversity

Our extensive data collection and the proposed DF-VAE method are designed to improve the *quality* of manipulated videos in DeeperForensics-1.0 dataset. In this section, we will mainly discuss the *scale* and *diversity* aspects.

We provide $10,000$ manipulated videos with 5 million frames. It is also an order of magnitude more than the previous datasets. We take $1,000$ refined YouTube videos collected by FaceForensics++ [23] as the target videos. Each face of our collected 100 identities is swapped onto 10 target videos, thus $1,000$ raw manipulated videos

TABLE 3.1: Seven types of distortions in DeeperForensics-1.0.

No.	Distortion Type
1	Color saturation change
2	Local block-wise distortion
3	Color contrast change
4	Gaussian blur
5	White Gaussian noise in color components
6	JPEG compression
7	Video compression rate change

are generated directly by DF-VAE in an end-to-end process. Thanks to the scalability and multimodality of DF-VAE, the time overhead of model training and data generation is reduced to 1/5 compared to the common Deepfakes methods, with no degradation in quality. Thus, a larger-scale dataset construction is possible.

To ensure *diversity*, we apply various perturbations to better simulate videos in real scenes. Specifically, as shown in Table 3.1, seven types of distortions defined in Image Quality Assessment (IQA) [153, 154] are included. Each of these distortions is divided into five intensity levels. We apply random-type distortions to the $1,000$ raw manipulated videos at five different intensity levels, producing a total of $5,000$ manipulated videos. Besides, an additional of $1,000$ robust manipulated videos are generated by adding random-type, random-level distortions to the $1,000$ raw manipulated videos. Moreover, in contrast to all the previous datasets, each sample of another $3,000$ manipulated videos in DeeperForensics-1.0 is subjected to a mixture of more than one distortion. The variability of perturbations improves the *diversity* of DeeperForensics-1.0 to better imitate the data distribution of real-world scenarios.

DeeperForensics-1.0 is a new *large-scale* dataset consisting of over $60,000$ videos with 17.6 million frames for real-world face forgery detection. *High-quality* source videos and manipulated videos constitute two main contributions of the dataset. The *diversity* of perturbations applying to the manipulated videos ensures the robustness of DeeperForensics-1.0 to simulate real scenes. The whole dataset is released, free to all research communities, for developing face forgery detection and more general human-face-related research.

TABLE 3.2: The percentage of user study ratings for the UADFV, DeepFake-TIMIT, Celeb-DF, FaceForensics++, Deep Fake Detection, DFDC, and DeeperForensics-1.0 dataset. A higher score means the users think the videos are more realistic.

Dataset	1	2	3	4	5	"real"
UADFV [76]	29.2	36.0	20.7	8.9	5.2	14.1%
DeepFake-TIMIT [78]	31.4	31.4	24.8	9.6	2.7	12.3%
Celeb-DF [80]	5.6	14.8	18.6	24.2	36.9	61.0%
FaceForensics++ [23]	46.8	31.4	13.4	4.4	4.0	8.4%
Deep Fake Detection [84]	26.0	28.0	24.1	11.5	10.3	21.9%
DFDC [85]	25.4	29.7	22.0	11.9	11.1	23.0%
DeeperForensics-1.0 (Ours)	4.3	8.9	22.6	29.8	34.3	**64.1%**

3.2.4 User Study

To examine the quality of DeeperForensics-1.0 dataset, we engage 100 professional participants, most of whom specialize in computer vision research. We believe these participants are qualified and well-trained in assessing realness of tempered videos. The user study is conducted on DeeperForensics-1.0 and six former datasets, *i.e.*, UADFV [76], DeepFake-TIMIT [78], Celeb-DF [80], FaceForensics++ [23], Deep Fake Detection [84], DFDC [85]. We randomly select 30 video clips from each of these datasets and prepare a platform for the participants to evaluate their realness. Similar to the user study of [155], the participants are asked to provide their feedbacks to the statement "The video clip looks real." and give scores at five levels (1-clearly disagree, 2-weakly disagree, 3-borderline, 4-weakly agree, 5-clearly agree. We assume that users who give a score of 4 or 5 think the video is "real"). The user study results are presented in Table 3.2. The quality of our dataset is appreciated by most of the participants. Compared to the previous datasets, DeeperForensics-1.0 achieves the highest realism rating. Although Celeb-DF [80] also gets very high realness scores, the scale of our dataset is much larger.

3.3 Video Forgery Detection Benchmark

Dataset split. In our benchmark, we exploit $1,000$ raw manipulated videos in Section 3.2.3 and $1,000$ YouTube videos from FaceForensics++ [23] as our *standard* set. The videos are split into training, validation, and test set with a ratio of $7:1:2$. The identities of the swapped faces may be duplicated because faces

of 100 invited actors are swapped onto 1,000 driving videos. To avoid data leak, we randomly choose unrepeated 70, 10, and 20 identities, and group all the videos according to the identities. Similar to [23], the test and training sets share a close distribution in our *standard* set.

Other experiments in our benchmark are conducted on different variants of the standard set. These variants share the same 1,000 driving videos with the standard set. We will detail them in Section 3.3.2. For a fair comparison, all the experiments are conducted in the same split setting.

Hidden test set. For real-world scenarios, some experiments conducted in previous works [23, 80] may not perform a convincing evaluation due to the huge biases caused by a close distribution between the training and the test sets. The aforementioned standard set has the same setting with these works. As a result, strong detection baselines obtain very high accuracy on the standard test set as demonstrated in Section 3.3.2. However, the ultimate goal of the face forensics dataset is to help detect forgery in real scenes. Even if the accuracy on the standard test set is high, the models may easily fail in real-world scenarios.

We argue that the test set of *real-world* face forgery detection *should not* share a close distribution with the training set. What we need is a test set that better simulates the real-world setting. We call it "hidden" test set. To better imitate fake videos in the real scene, the hidden test set should satisfy three factors: 1) *Multiple sources.* Fake videos in-the-wild should be manipulated by different unknown methods. 2) *High quality.* Threatening fake videos should have high quality to fool human eyes. 3) *Diverse distortions.* Different perturbations should be taken into consideration.

Thus, in our initial benchmark, we introduce a challenging hidden test set with 400 carefully selected videos. First, we collect fake videos generated by several unknown face swapping methods to ensure multiple sources. Then, we obscure all selected videos multiple times with diverse hidden distortions that are commonly seen in real scenes. Finally, we only select videos that can fool at least 50 out of 100 human observers in a user study. The ground truth labels are hidden and are used on our host server to evaluate the accuracy of detection models. Besides, the hidden test set will be enlarged constantly to get future versions along with development of

TABLE 3.3: The binary detection accuracy of the baselines on the hidden test set when trained on four manipulated methods in FaceForensics++ (FF++): DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT), and on DeeperForensics-1.0 standard training set without distortions.

Train Test (acc)	FF++ DF hidden	FF++ F2F hidden	FF++ FS hidden	FF++ NT hidden	**DeeperForensics-1.0** **hidden**
C3D [100]	57.50	57.75	52.13	58.25	**74.75**
TSN [101]	57.63	57.25	53.50	57.38	**77.00**
I3D [97]	56.63	58.38	54.63	63.63	**79.25**
ResNet+LSTM [98, 99]	57.38	56.13	54.88	59.50	**78.25**
XceptionNet [89]	57.38	58.75	54.75	57.38	**77.00**

Deepfakes technology. Fake videos manipulated by future face swapping methods will be included as long as they can pass the human test supported by us.

3.3.1 Baselines

Existing studies [23, 80] primarily provide image-level face forgery detection benchmark. However, fake videos in-the-wild are much more menacing than manipulated images. We propose to conduct evaluation mainly based on video classification methods for two reasons. First, image-level face forgery detection methods do not consider any temporal information – an important cue for video-based tasks. Second, image-level methods have been widely studied. We only choose one image-level method, XceptionNet [89], which achieves the best performance in [23], as one part of our benchmark for reference. The other four video-based baselines are C3D [100], TSN [101], I3D [97], and ResNet+LSTM [98, 99], all of which have achieved promising results in video classification tasks.

3.3.2 Results and Analysis

Owing to the goal of detecting fakes in real-world scenarios, we mainly explore how common distortions appearing in real scenes affect the model performance. Accuracies of face forgery detection on the standard test set and the introduced hidden test set are evaluated under various settings.

Evaluation of effectiveness of DeeperForensics-1.0. For a fair comparison, we evaluate DeeperForensics-1.0 and the state-of-the-art FaceForensics++ [23] dataset because they use the same driving videos. In this setting, we use $1,000$ raw

TABLE 3.4: The binary detection accuracy of the baselines when trained and tested on DeeperForensics-1.0 dataset with different distortion perturbations. We analyze different training and testing settings on the standard set without distortions (std), the standard set with single-level distortions (std/sing), and the standard set with random-level distortions (std/rand).

Train Test (acc)	**std** **std**	std std/sing	std std/rand	std/sing std/sing	std/rand std/rand	std/sing std/rand	std/rand std/ sing
C3D [100]	**98.50**	87.63	92.38	95.38	96.63	96.75	94.00
TSN [101]	**99.25**	91.50	95.00	98.25	98.88	98.12	99.12
I3D [97]	**100.00**	90.75	96.88	99.50	99.63	99.63	98.00
ResNet+LSTM [98, 99]	**100.00**	90.63	97.13	100.00	98.63	100.00	97.25
XceptionNet [89]	**100.00**	88.38	94.75	99.63	99.63	99.75	99.00

manipulated videos without distortions in the standard set of DeeperForensics-1.0. For FaceForensics++, the same split is applied to its *four* subsets. All the models are tested on the hidden test set (see Table 3.3).

The baselines trained on the standard training set of DeeperForensics-1.0 achieve much better performance on the hidden test set than all the *four* subsets of FaceForensics++. This proves the higher *quality* of DeeperForensics-1.0 over prior works, making it more useful for real-world face forgery detection. In Table 3.3, I3D [97] obtains the best performance on the hidden test set when trained on the standard training set. We conjecture that the temporal discontinuity of fake videos leads to higher accuracy by this video-level forgery detection method.

Evaluation of dataset perturbations. We study the effect of perturbations towards the forgery detection model performance. In contrast to prior work [23], we try to evaluate the baseline accuracies when applying different distortions to the training and the test sets, in order to explore the function of perturbations in face forensics dataset.

In this setting, we conduct all the experiments on DeeperForensics-1.0 dataset with high diversity of perturbations. We use $1,000$ manipulated videos in the standard set (std), $1,000$ manipulated videos with single-level (level-5), random-type distortions (std/sing), $1,000$ manipulated videos with random-level, random-type distortions (std/rand). The data split is the same as that of the standard set with a ratio of $7:1:2$.

In Column 2 of Table 3.4, we find the accuracy is nearly 100% when the models are trained and tested on the standard set. This is reasonable because the strong baselines perform very well in a clean dataset with the same distribution. In

TABLE 3.5: The binary detection accuracy of the baselines on the hidden test set when trained on DeeperForensics-1.0 dataset with the standard set without distortions (std), combination of std and the standard set with single-level distortions (std+std/sing), combination of std and the standard set with random-level distortions (std+std/rand), combination of std and the standard set with the mixed distortions(std+std/mix).

Train Test (acc)	std hidden	std+std/sing hidden	std+std/rand hidden	**std+std/mix hidden**
C3D [100]	74.75	78.25	78.13	**78.88**
TSN [101]	77.00	78.75	79.50	**79.50**
I3D [97]	79.25	80.13	80.13	**80.13**
ResNet+LSTM [98, 99]	78.25	80.25	79.50	**80.25**
XceptionNet [89]	77.00	79.75	79.75	**79.88**

Columns 3 and 4, the accuracy decrease compared to Column 2, when we choose std/sing and std/rand as the test set. Most of the video-level methods except C3D [100] are more robust to perturbations on test set than XceptionNet [89]. This setting is very common because different distributions of the training and the test sets lead to decrease in model accuracies. Hence, the lack of perturbations in the face forensics dataset cutbacks the model performance for real-world face forgery detection with even more complex data distribution.

When we apply corresponding distortions to the training and test sets, the accuracy will increase (Column 5 and 6 in Table 3.4) compared to Column 3 and 4. However, this setting is impractical because the distributions of the training and test sets are still the same. We should augment the test set to better simulate the real-world distribution. Thus, some evaluation settings in previous works [23, 80] are unreasonable. If we swap the training set and the test set of std/sing and std/rand to further randomize the condition, results shown in Column 7 and 8 indicate that the accuracy remains high. This evaluation setting shows the possibility that with the same generation method, exerting appropriate distortions to the training set can make face forgery detection models more robust to real-world perturbations.

Evaluation of variants of training set for real-world face forgery detection. We have conducted several experiments for evaluations of possible perturbations. Nevertheless, the case is more complex in real scenes because no information about the fake videos is available. The video may be subjected to more than one type and diverse levels of distortions. In addition to distortions, the method manipulating the faces is unknown.

From the evaluation of perturbations, we find the possibility of augmenting the training set to improve detection model performance. Thus, we further evaluate baseline performance on the hidden test set by devising some variants of the training set. We perform experiments on DeeperForensics-1.0. In this setting, other than std, std/sing, and std/rand, we use additional $1,000$ manipulated videos, each of which is subjected to a mixture of three random-level, random-type distortions (std/mix). We combine std with std/sing, std/rand, and std/mix, respectively, yielding three new training sets (with the same data split as the former settings).

Column 2 in Table 3.5 shows the low accuracy when the models trained on std and tested on the hidden test set (same as Column 6 in Table 3.3). Columns 3 and 4 indicate that the accuracy of all the baseline models increase when trained on std+std/sing and std+std/rand. The accuracy of I3D [97] and ResNet+LSTM [98, 99], are over 80% in some cases. In a more complex setting, when the models are trained on std+std/mix, Column 5 shows the accuracy of all the detection baselines further increase.

The results suggest that designing suitable training set variants has the potential to help increase face forgery detection accuracy, and applying various distortions to ensure the *diversity* of DeeperForensics-1.0 is necessary. In addition, compared to image-level method, video-level face forgery detection methods have more potential capabilities to crack real-world fake videos as shown in Table 3.5.

Although the accuracy on the challenging hidden test set is still not very high, we provide two initial directions for future real-world face forgery detection research: 1) Improving the source data collection and generation method to ensure the *quality* of the training set; 2) Augmenting the training set by various distortions to ensure its *diversity*. We welcome researchers to make our benchmark more comprehensive.

3.4 Discussion

In this work, we propose a new large-scale dataset named DeeperForensics-1.0 to facilitate the research of face forgery detection towards *real-world* scenarios. We make several efforts to ensure *good quality*, *large scale*, and *high diversity* of this dataset. Based on the dataset, we further benchmark existing representative

forgery detection methods, offering insights into the current status and future strategy in face forgery detection. Several topics can be considered as future works. 1) We will continue to collect more source and target videos to further expand DeeperForensics. 2) We plan to invite interested researchers for contributing their video falsification methods to enlarge our hidden test set, as long as the fakes can pass the human test supported by us. 3) A better evaluation metric for face forgery detection methods is also an interesting research topic.

Chapter 4

TSIT: A Simple and Versatile Framework for Image-to-Image Translation

4.1 Introduction

After securing the potential countermeasures against the negative impact of generated data (Chapter 3), we wish to dive more deeply into the potential of our research in many other practical applications. A versatile framework is useful as it circumvents cumbersome modifications when applied to different tasks, which is meaningful in practice. We then shift our attention to another important problem in image generation, *i.e.*, image-to-image translation [3], which aims at translating one image representation to another. Recent advances [2, 10, 12, 25, 26], especially GANs [10], have made remarkable success in various image-to-image translation tasks. Previous studies usually present specialized solutions for a specific form of application, ranging from arbitrary style transfer [14, 19, 20, 27, 61, 64, 104] under the unsupervised setting, to semantic image synthesis [3, 32–35, 65] in the supervised setting.

In this chapter, we are interested in devising a general and unified framework that is applicable to different image-to-image translation tasks with a negligible sacrifice

* The work in this chapter has been published in [21].

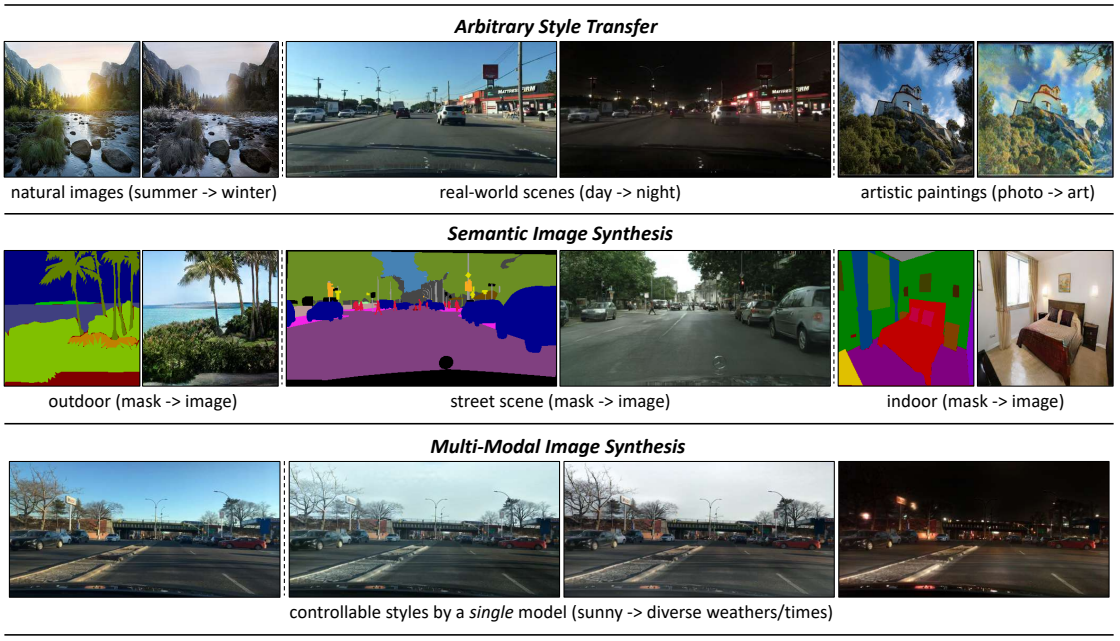

FIGURE 4.1: Our TSIT framework is simple and versatile for various image-to-image translation tasks. For unsupervised arbitrary style transfer, diverse scenarios (*e.g.*, natural images, real-world scenes, artistic paintings) can be handled. For supervised semantic image synthesis, our method is robust to different scenes (*e.g.*, outdoor, street scene, indoor). Multi-modal image synthesis is feasible by a *single* model with controllable styles.

of synthesis quality. This is *non-trivial* given the different natures of different tasks. For instance, in certain conditional image synthesis tasks (*e.g.*, arbitrary style transfer), paired data are usually not available. Under this unsupervised setting, translation task demands additional constraints on cycle consistency [14, 19, 27, 28], semantic features [29], pixel gradients [30], or pixel values [31]. In semantic image synthesis (*i.e.*, translation from segmentation labels to images), training pairs are available. This task is more data-dependent and typically needs losses to minimize per-pixel distance between the generated sample and ground truth. In addition, specialized structures [32–35] are required to maintain spatial coherence and resolution. Due to the different needs, existing methods exploit their own specially designed components. It is difficult to cross-use these components or integrate them into a unified framework.

To address the aforementioned challenges, we propose a Two-Stream Image-to-image Translation (TSIT) framework, which is *versatile* for various image-to-image translation tasks (see Figure 4.1). The framework is simple as it is based purely on feature transformation. Unlike previous approaches [34, 104] that only consider

either semantic structure or style representation, we factorize *both* the structure and style in multi-scale *feature levels* via a symmetrical *two-stream* network. The two streams jointly influence the new image generation in a coarse-to-fine manner via a consistent feature transformation scheme. Specifically, the content spatial structure is preserved by an element-wise feature adaptive denormalization (FADE) from the content stream, while the style information is exerted by feature adaptive instance normalization (FAdaIN) from the style stream. Standard loss functions such as adversarial loss and perceptual loss are used, without additional constraints like cycle consistency. The pipeline is applicable to both unsupervised and supervised settings, easing the preparation of data.

The **contributions** of our work are summarized as follows. We propose TSIT, a simple and versatile framework, which is effective for various image-to-image translation tasks. Despite the succinct design, our network is readily adaptable to various tasks and achieves compelling results. The good performance is achieved by 1) *multi-scale* feature normalization (FADE and FAdaIN) scheme that captures *coarse-to-fine* structure and style information, and 2) a *two-stream* network design that integrates *both* content and style effectively, reducing artifacts and making multi-modal image synthesis possible (see Figure 4.1). In comparison to several state-of-the-art task-specific baselines [20, 32–35, 64, 65], our method achieves comparable or even better results in both perceptual quality and quantitative evaluations.

4.2 Methodology

We consider three key requirements in formulating a robust and scalable method to link various tasks: 1) *Both* semantic structure information and style representation should be considered and fused adaptively. 2) The content and style information should be learned by networks in *feature level* instead of in image level to fit the nature of diverse semantic tasks. 3) The network structure and loss functions should be *simple* for easy training without additional constraints.

Based on the aforementioned considerations, we design a TSIT framework (see Figure 4.2). We will detail our method in this section, including the network

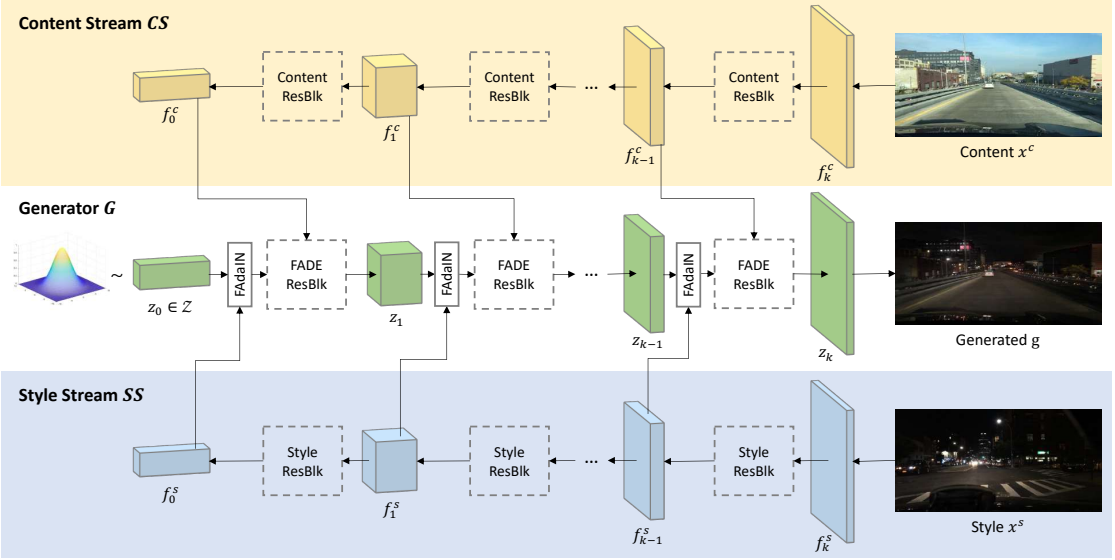

FIGURE 4.2: The proposed Two-Stream Image-to-image Translation (TSIT) framework. The multi-scale patch-based discriminators are omitted. A Gaussian noise map is taken as the latent input for the generator. The feature representations of the content and style images are extracted by the corresponding streams for multi-scale feature transformations. The symmetrical networks fuses semantic structure and style representation in an end-to-end training. Submodules of our network are shown in Figure 4.3.

structure (Section 4.2.1), the feature transformation scheme (Section 4.2.2), and the objective functions (Section 4.2.3).

4.2.1 Network Structure

As illustrated in Figure 4.2, TSIT consists of four components: content stream, style stream, generator, and discriminators (omitted in Figure 4.2). The first three main components are fully convolutional and symmetrically designed. The details of the submodules, including content/style residual block, FADE residual block, FADE module in the FADE residual block, are as shown in Figure 4.3. We will discuss them separately in this section.

Content/style stream. Unlike the traditional conditional GAN [12], we place the two-stream networks, *i.e.*, content stream and style stream, on each side of the generator (see Figure 4.2). These two streams are symmetrical with the same network structure, aiming at extracting corresponding feature representations in different levels. We construct content/style stream based on standard residual

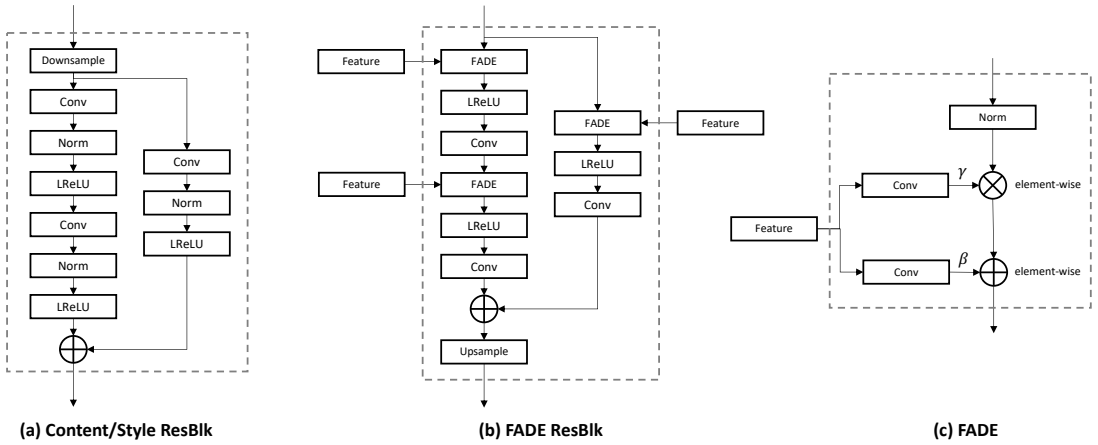

FIGURE 4.3: Submodules of our framework. (a) is a content/style residual block in the symmetrical content/style streams. (b) is a FADE residual block in the generator. (c) is a FADE module in the FADE residual block. It performs *element-wise* denormalization by modulating the normalized activation using a learned affine transformation defined by the modulation parameters γ and β.

blocks [98]. We call them content/style residual blocks. As shown in Fig 4.3 (a), each block has three convolutional layers, one of which is designed for the learned skip connection. The activation function is Leaky ReLU. The function of content/style stream is to extract features and feed them to the corresponding feature transformation layers in the generator. Multi-scale content/style representation in *feature levels* can be learned by content/style stream, adaptively fitting different feature transformations.

Generator. The generator has a completely inverse structure *w.r.t.* the content/style stream. This is intentionally designed to consistently match the level of semantic abstraction at different feature scales. A noise map is sampled from a Gaussian distribution as the latent input, and the feature maps from corresponding layers in content/style stream are taken as multi-scale feature inputs. The proposed feature transformations are implemented by a FADE residual block (Figure 4.3 (b)) and a FAdaIN module. In the FADE residual block, we use an inverse architecture *w.r.t.* the content/style residual block and replace the batch normalization [156] layer with the FADE module (Figure 4.3 (c)). The FADE module performs *element-wise* denormalization by modulating the normalized activation using a learned affine transformation defined by the modulation parameters γ and β. The FAdaIN module is used to exert style information through feature adaptive instance normalization. More discussions are given in Section 4.2.2.

The entire image generation process is performed in a coarse-to-fine manner. In particular, multi-scale content/style features are injected to refine the generated image constantly from high-level latent code to low-level image representation. Semantic structure and style information are learnable and effectively fused in an end-to-end training.

Discriminators. We exploit the standard multi-scale patch-based discriminators (omitted in Figure 4.2) in [33, 34]. Three regular discriminators with an identical architecture are included to discriminate images at different scales. Despite the same structure, patch-based training allows the discriminator operating at the coarsest scale to have the largest receptive field, capturing global information of the image. Whereas the one operating at the finest scale has the smallest receptive field, making the generator produce better details. Multi-scale patch-based discriminators further improve the robustness of our method for image-to-image translation tasks in different resolutions. Besides, the discriminators also serve as feature extractors for the generator to optimize the feature matching loss.

4.2.2 Feature Transformation

We propose a new feature transformation scheme, considering *both* semantic structure information and style representation, and fusing them adaptively. Let x^c be the content image and x^s be the style image. CS, SS, G, D denote content stream, style stream, generator, and discriminators, respectively. Sampled from a Gaussian distribution, $z_0 \in \mathbb{Z}$ is a noise map as the latent input for the generator (Figure 4.2). Let $z_i \in \{z_0, z_1, z_2, ..., z_k\}$ be the feature map after i-th residual block in the generator, with k denoting the the total number of residual blocks (*i.e.*, the upsampling times in the generator). Let $f_i^c \in \{f_0^c, f_1^c, f_2^c, ..., f_k^c\}$ represent the corresponding feature representations extracted by the content stream (Figure 4.2), $f_i^s \in \{f_0^s, f_1^s, f_2^s, ..., f_k^s\}$ with the similar meaning in the style stream.

Feature adaptive denormalization (FADE). Our method is inspired by spatially adaptive denormalization (SPADE) [34]. Different from SPADE that resizes a semantic mask as its input, we generalize the input to multi-scale *feature representation* f_i^c of the content image x^c. In this way, we fully exploit semantic information captured by the content stream CS.

Formally, we define N as the batch size, L_i as the number of feature map channels in each layer. H_i and W_i are height and width, respectively. We first apply batch normalization [156] to normalize the generator feature map z_i in a channel-wise manner. Then, we modulate the normalized feature by using the learned parameters scale γ_i and bias β_i. The denormalized activation ($n \in N$, $l \in L_i$, $h \in H_i$, $w \in W_i$) is:

$$\gamma_i^{l,h,w} \cdot \frac{z_i^{n,l,h,w} - \mu_i^l}{\sigma_i^l} + \beta_i^{l,h,w}, \tag{4.1}$$

where μ_i^l and σ_i^l are the mean and standard deviation, respectively, of the generator feature map z_i before the batch normalization [156] in channel l:

$$\mu_i^l = \frac{1}{NH_iW_i} \sum_{n,h,w} z_i^{n,l,h,w}, \tag{4.2}$$

$$\sigma_i^l = \sqrt{\frac{1}{NH_iW_i} \sum_{n,h,w} \left(z_i^{n,l,h,w}\right)^2 - \left(\mu_i^l\right)^2}. \tag{4.3}$$

The denormalization operation is *element-wise*, and the parameters $\gamma_i^{l,h,w}$ and $\beta_i^{l,h,w}$ are learned by one-layer convolutions from f_i^c in the FADE module (see Figure 4.3 (c)). Compared to previous conditional normalization methods [34, 104, 107], FADE experiences more perceptible influence from coarse-to-fine feature representations, thus it can better preserve semantic structure information.

Feature adaptive instance normalization (FAdaIN). To better fuse style representation, we introduce another feature transformation, named feature adaptive instance normalization (FAdaIN). This method is inspired by adaptive instance normalization (AdaIN) [104], with a generalization to enable the style stream SS to learn multi-scale *feature-level* style representation f_i^s of the style image x^s more effectively.

We use the same notation z_i to represent the feature map after i-th FADE residual block in the generator. FAdaIN adaptively computes the affine parameters from the corresponding style feature f_i^s with the same scale from SS:

$$\text{FAdaIN}\left(z_i, f_i^s\right) = \sigma\left(f_i^s\right)\left(\frac{z_i - \mu\left(z_i\right)}{\sigma\left(z_i\right)}\right) + \mu\left(f_i^s\right), \tag{4.4}$$

where $\mu(z_i)$ and $\sigma(z_i)$ are the mean and standard deviation, respectively, of z_i.

Exploiting FAdaIN, coarse-to-fine style features at different layers can be fused adaptively with the corresponding semantic structure features learned by FADE, allowing our framework to be trained end-to-end and versatile to different tasks. Furthermore, owing to the effectiveness of FAdaIN in capturing multi-scale style feature representations, multi-modal image synthesis is made possible with arbitrary style control.

4.2.3 Objective

We use standard losses in our objective function. Following [34, 35], we adopt a hinge loss term [148, 157, 158] as our adversarial loss. For the generator, we apply hinge-based adversarial loss, perceptual loss [66], and feature matching loss [33]. For the multi-scale discriminators, only hinge-based adversarial loss is used to distinguish whether the image is real or fake. The generator and discriminator are trained alternately to play a min-max game. The generator loss $\mathcal{L}_G$ and the discriminator loss $\mathcal{L}_D$ can be written as:

$$\mathcal{L}_G = -\mathbb{E}\left[D\left(g\right)\right] + \lambda_P \mathcal{L}_P\left(g, x^c\right) + \lambda_{FM} \mathcal{L}_{FM}\left(g, x^s\right), \tag{4.5}$$

$$\mathcal{L}_D = -\mathbb{E}\left[\min\left(-1 + D\left(x^s\right), 0\right)\right] - \mathbb{E}\left[\min\left(-1 - D\left(g\right), 0\right)\right], \tag{4.6}$$

where $g = G\left(z_0, x^c, x^s\right)$ denotes the generated image, z_0, x^c, x^s denote the input noise map in latent space, the content image, and the style image, respectively. $\mathcal{L}_P$ is the perceptual loss [66] that minimizes the difference between the feature representations extracted by VGG-19 [66] network. $\mathcal{L}_{FM}$ is the feature matching loss [33] that matches the intermediate features at different layers of multi-scale discriminators. λ_P and λ_{FM} are the corresponding weights.

The simple objective functions make our framework stable and easy to train. Thanks to the two-stream network, the typical KL loss [2] for multi-modal image synthesis becomes optional. Despite the simplicity, TSIT is highly versatile for various image-to-image translation tasks.

4.3 Experiments

4.3.1 Settings

Implementation details. We use Adam [159] optimizer and set $\beta_1 = 0$, $\beta_2 = 0.9$. Two time-scale update rule [8] is applied, where the learning rates for the generator (including two streams) and the discriminators are 0.0001 and 0.0004, respectively. We exploit Spectral Norm [148] for all layers in our network. We adopt SyncBN and IN [160] for the generator and the multi-scale discriminators, respectively. For the perceptual loss [66], we use the feature maps of `relu1_1`, `relu2_1`, `relu3_1`, `relu4_1`, `relu5_1` layers from a pretrained VGG-19 [161] model, with the weights [1/32, 1/16, 1/8, 1/4, 1]. For the feature matching loss [33], we select features of three layers from the discriminator at each scale. All the experiments are conducted on NVIDIA Tesla V100 GPUs.

Applications. The proposed framework is versatile for various image-to-image translation tasks. We consider three representative applications of conditional image synthesis: arbitrary style transfer (unsupervised), semantic image synthesis (supervised), and multi-modal image synthesis (enriching generation diversity).

Datasets. For arbitrary style transfer, we consider diverse scenarios. We use Yosemite summer $\rightarrow$ winter dataset (natural images) provided by [14]. We classify BDD100K [162] (real-world scenes) into different times and perform day $\rightarrow$ night translation. Besides, we use Photo $\rightarrow$ art dataset (artistic paintings) in [14]. For semantic image synthesis, we select several challenging datasets (*i.e.*, Cityscapes [163] and ADE20K [164]). For multi-modal image synthesis, we further classify BDD100K [162] into different time and weather conditions, and perform controllable time and weather translation.

Evaluation metrics. Besides comparing perceptual quality, we employ the standard evaluation protocol in prior works [5, 13, 20, 34, 35] for quantitative evaluation. For arbitrary style transfer, we apply Fréchet Inception Distance (FID, evaluating similarity of distribution between the generated images and the real images, lower is better) [8] and Inception Score (IS, considering clarity and diversity, higher is better) [51]. For semantic image synthesis, we strictly follow [34, 35], adopting FID [8] and segmentation accuracy (mean Intersection-over-Union (mIoU) and pixel

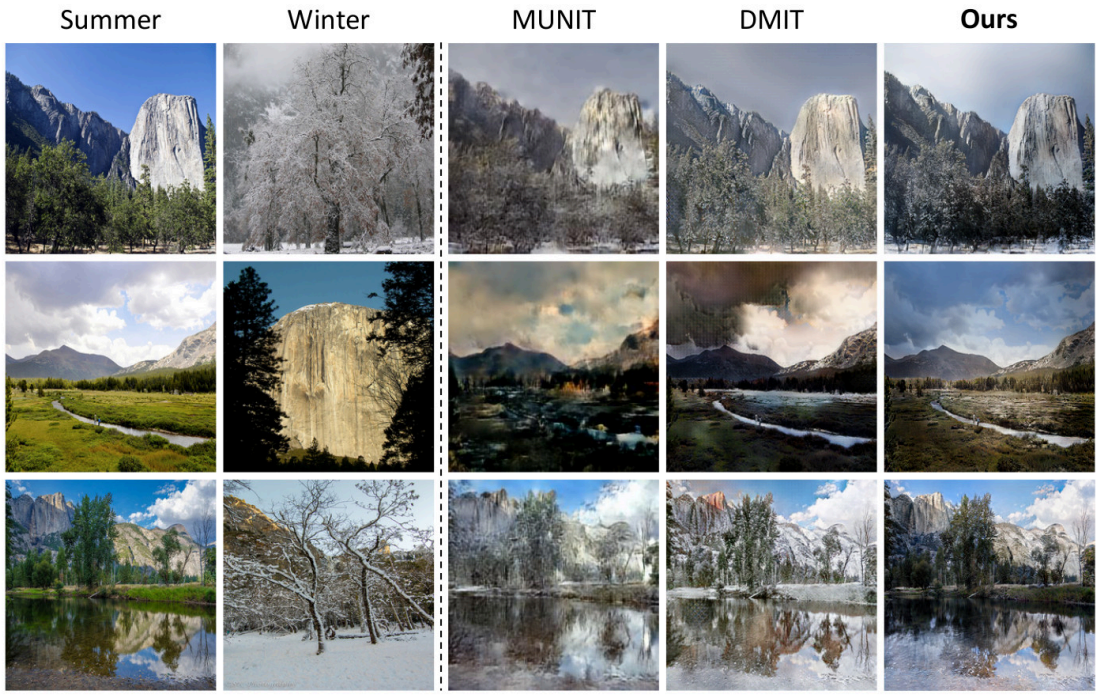

FIGURE 4.4: **Yosemite summer → winter** season transfer results compared to baselines.

accuracy (accu)). The segmentation models are: DRN-D-105 [165] for Cityscapes [163], and UperNet101 [166] for ADE20K [164].

Baselines. We compare our method with several state-of-the-art task-specific baselines. For a fair comparison, we mainly employ GAN-based methods. In the unsupervised setting, MUNIT [20] and DMIT [64] are included, with the strong ability to capture the multi-modal nature of images while keeping quality. In the supervised setting, we compare against CRN [32], SIMS [65], pix2pixHD [33], SPADE [34], and CC-FPSE [35].

4.3.2 Results and Analysis

Arbitrary style transfer. The results of *Yosemite summer → winter season transfer* are shown in Figure 4.4. Baselines [20, 64] tend to impose the color of the style image (winter) to the whole content image (summer). Besides, MUNIT sometimes introduces unnecessary artistic effects, and DMIT generates some grid-like artifacts. In comparison, our generated results are clearer and more semantics-aware spatially. The results of *BDD100K day → night time translation* are shown

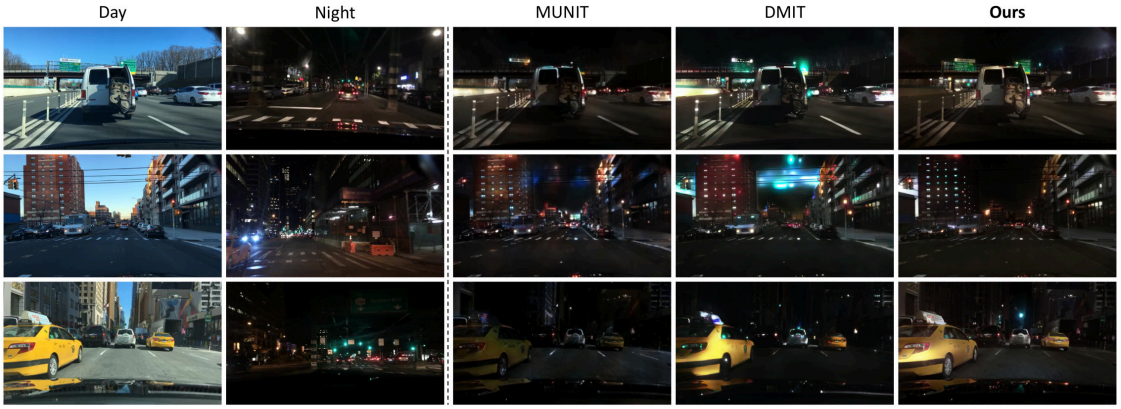

FIGURE 4.5: **BDD100K day → night** time translation results compared to baselines.

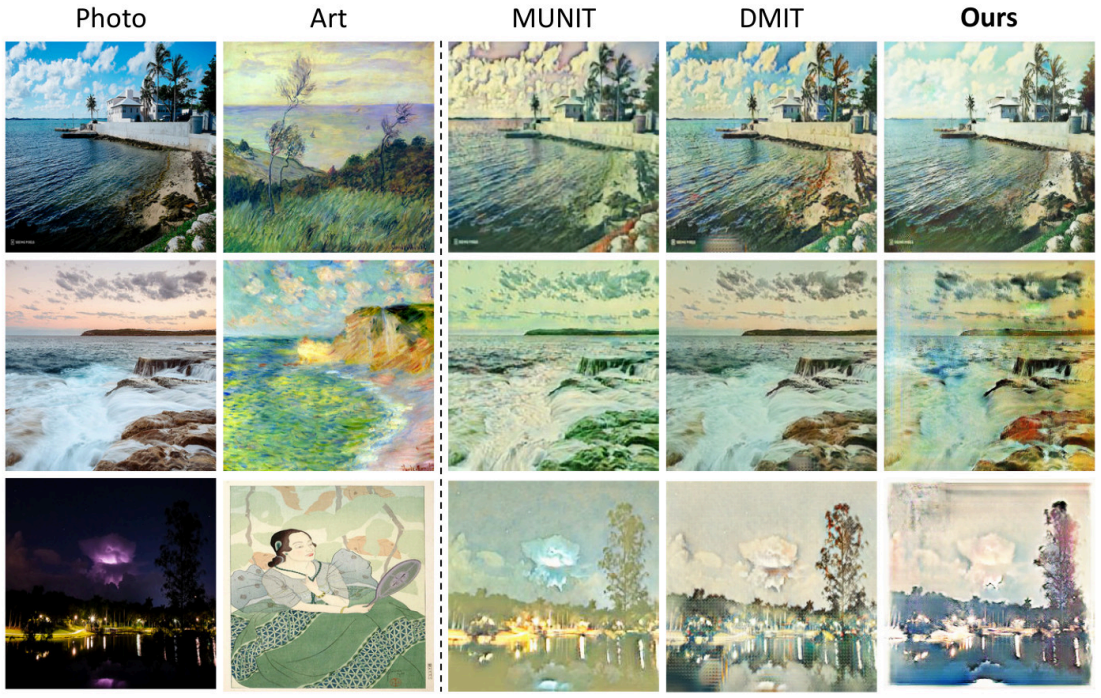

FIGURE 4.6: **Photo → art** style transfer results compared to baselines.

TABLE 4.1: The FID and IS scores of our method compared to state-of-the-art methods in arbitrary style transfer tasks. A lower FID and a higher IS indicate better performance.

	summer → winter		day → night		photo → art	
Methods	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑
MUNIT [20]	118.225	2.537	110.011	2.185	167.314	3.961
DMIT [64]	87.969	2.884	83.898	2.156	166.933	3.871
Ours	**80.138**	**2.996**	**79.697**	**2.203**	**165.561**	**4.020**

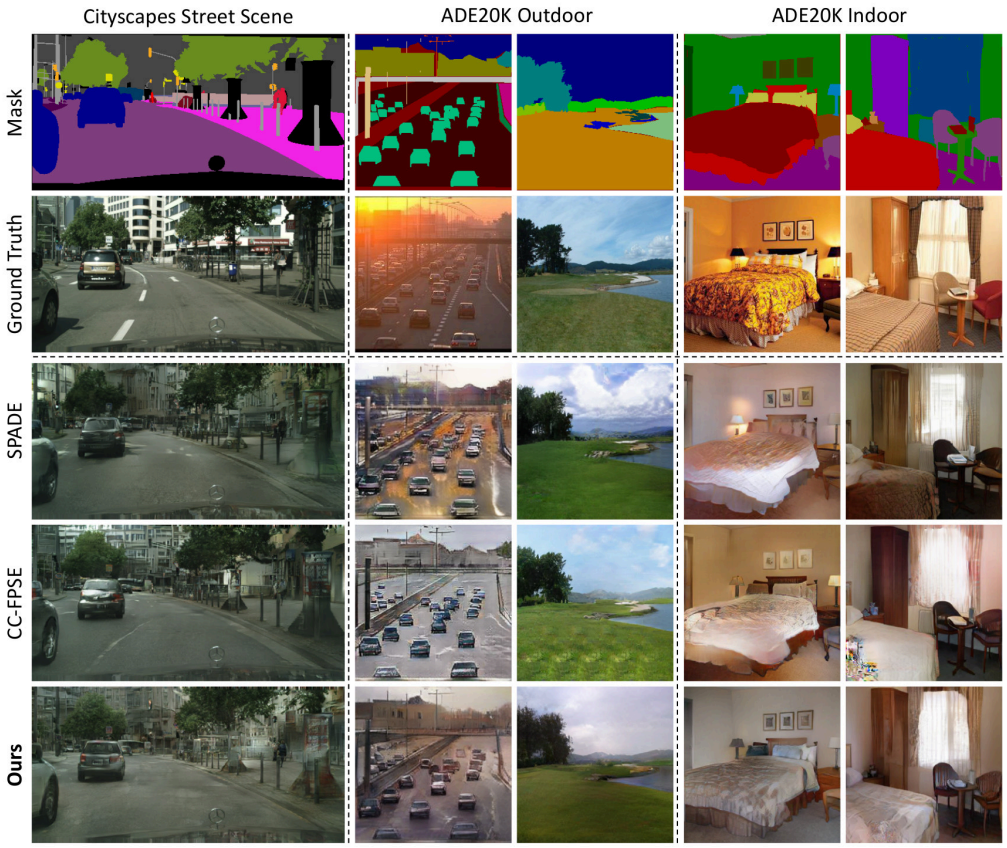

FIGURE 4.7: **Semantic image synthesis** results compared to baselines.

in Figure 4.5. Some objects (*e.g.*, road sign, car) generated by MUNIT are too dark, and the whole image tends to have some unnatural colors. DMIT introduces obvious artifacts to the car or sky. In contrast, our method produces more photo-realistic samples in this task. In *photo* → *art style transfer*, we choose some hard cases to make a clear comparison (see Figure 4.6) due to the very strong ability of all the methods in this task. Our method can transfer the styles well while effectively keeping the content structure. MUNIT tends to impose a homogeneous color to the image. Although DMIT achieves slightly better stylization than our method in certain cases (in Row 3 of Figure 4.6), it also brings some grid-like distortions.

The quantitative evaluation results are shown in Table 4.1. Our approach achieves better performance than baselines [20, 64] in all the tasks. We also note that the gap is relatively small in photo → art style transfer, in line with the close qualitative performance in this task (see Figure 4.6).

Semantic image synthesis. We choose two state-of-the-art baselines, SPADE [34] and CC-FPSE [35], to show some qualitative comparison results of semantic image synthesis (Figure 4.7). Our method demonstrates better perceptual quality

TABLE 4.2: The mIoU, pixel accuracy (accu), and FID scores of our method compared to state-of-the-art methods in semantic image synthesis tasks. A higher mIoU, a higher pixel accuracy (accu), and a lower FID indicate better performance.

	Cityscapes			ADE20K		
Methods	mIoU ↑	accu ↑	FID ↓	mIoU ↑	accu ↑	FID ↓
CRN [32]	52.4	77.1	104.7	22.4	68.8	73.3
SIMS [65]	47.2	75.5	**49.7**	N/A	N/A	N/A
pix2pixHD [33]	58.3	81.4	95.0	20.3	69.2	81.8
SPADE [34]	62.3	81.9	71.8	38.5	79.9	33.9
CC-FPSE [35]	65.5	82.3	54.3	**43.7**	**82.9**	31.7
Ours	**65.9**	**82.7***	59.2	38.6	80.8	**31.6**

than these task-specific baselines. In street scene (Column 1), our method generates better details on key objects (car, pedestrian). In road scene (Column 2), SPADE generates atypical colors on the roads, while CC-FPSE produces unnatural edges on the cars, hardly fitting the background (road). For outdoor natural images (Column 3), all the methods share a similar generation quality. Our method is slightly better due to less distortions on the grass. In indoor scene (Column 4 and 5), SPADE and CC-FPSE produce obvious artifacts in some cases (Column 5). In contrast, our method is more robust to diverse scenarios.

The quantitative evaluation results are shown in Table 4.2 (the values used for comparison are taken from [34, 35]). The proposed method achieves comparable performance with the very strong specialized methods [32–35, 65] for semantic image synthesis. Note that SIMS [65] yields the best FID score but poor segmentation performance on Cityscapes, because it stitches image patches from a memory bank of training set while not keeping the exactly consistent position in the synthesized image. Our approach achieves state-of-the-art segmentation performance on Cityscapes and the best FID score on ADE20K, suggesting its robustness to fit the nature of different image-to-image translation tasks.

Multi-modal image synthesis. We perform multi-modal image synthesis for time and weather image-to-image translation (see Figure 4.8) on BDD100K [162].

* The value differs from the earlier version of this work [167]. The official code of DRN [165] does not provide the implementation of the "accu" metric. The new accu value 82.7% (still the best among the compared methods) is obtained by including 255-labeled pixels, consistent with [34, 35]. The previously reported accu 94.4% omits 255-labeled pixels, which may be more reasonable due to its consistency with the training of the segmentation model and the calculation of mIoU.

FIGURE 4.8: **BDD100K multi-modal image synthesis** results for different time and weather translation by a *single* model.

FIGURE 4.9: **Cross validation** of ineffectiveness of task-specific methods in inverse settings.

Training only a *single* model, we translate the images of weather *sunny* to different times and weathers (*i.e.*, *night*, *snowy*, *cloudy*, *rainy*). Our method effectively adapts to different style control and keeps photorealistic generation quality. Although the weather *snowy* is not very obvious in BDD100K [162], our approach successfully introduces some snow-like effects on trees and grounds (Column 2).

Cross validation. We also conduct experiments to evaluate the performance of existing specialized methods in inverse settings (*i.e.*, using unsupervised methods to do semantic image synthesis / using supervised methods to perform arbitrary style transfer). We selected two representative methods, MUNIT [20] and SPADE [34]. Without modifying the architecture, we tuned the loss weights and tried to get the best generation results. To ensure a fair comparison, we also tried to

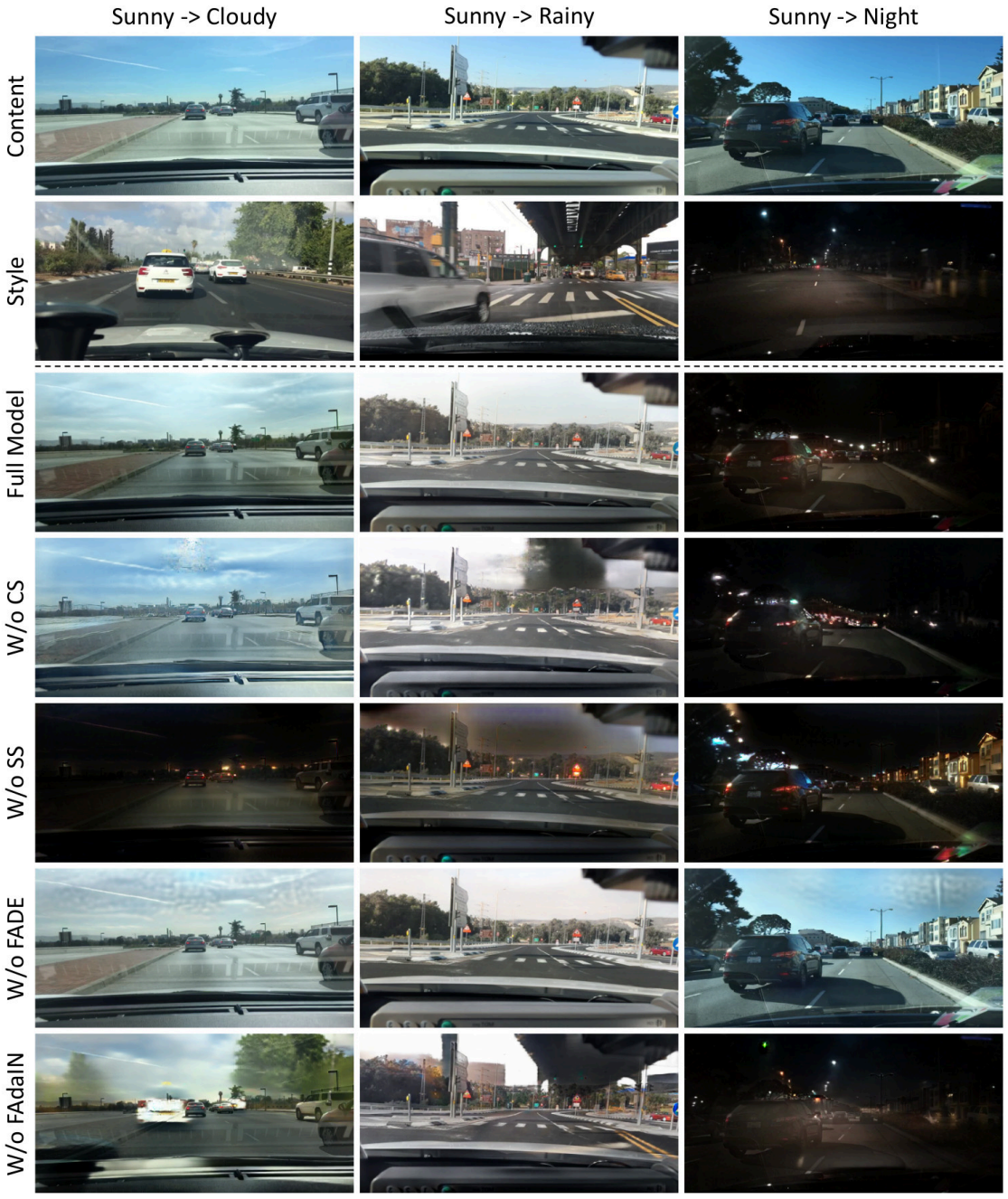

FIGURE 4.10: **Ablation studies** of key modules (*i.e.*, content stream (CS), style stream(SS)) and feature transformations in the multi-modal image synthesis task.

compute perceptual loss with the content (day) image for SPADE to match the setting of TSIT. Representative results of cross validation are shown in Figure 4.9. The proposed method shows much better results than baseline methods. MUNIT fails to adapt to semantic image synthesis. SPADE loses details of key objects and introduces very strong artifacts despite translating the color correctly.

Ablation studies. We present ablation studies of key modules (*i.e.*, content stream (CS), style stream(SS)) and the proposed feature transformations (see Figure 4.10). We perform multi-modal image synthesis to show the effectiveness of different components. Our full model generates high-quality results (Row 3). When we directly inject the resized content image without CS, the semantic structure information becomes weak, leading to several artifacts in the sky (Row 4). Without SS, the model cannot perform multi-modal image synthesis at all (Row 5). The style representation is dominated by the night style. When we concatenate the feature maps of CS with the ones of the generator instead of using FADE, the concatenation introduces too much content information, leading to several failure cases (*e.g.*, *sunny* $\rightarrow$ *night* in Row 6). If we discard FAdaIN by concatenating the feature maps of SS with the ones of the generator, the style becomes too strong, causing serious style regionalization problem (Row 7).

4.4 Limitations and Failure Cases

Notwithstanding the promising results, the introduced TSIT framework can still be improved in a few aspects. First, the overall quality of synthesis may be further improved. Several observable artifacts might exist on the generated images. In particular (see Figure 4.11), unnatural regional stylization (Columns 1-3) and spot artifacts (Column 4) may occur in arbitrary style transfer. Second, the style representation is sometimes fused globally and randomly, weakening the semantic-aware stylization. We attribute this issue to the lack of semantic guidance in the style stream with only random initialization. Third, the TSIT network is a bit heavy and slow, despite its versatility for various tasks. The heavy structure still limits its scalability to many real-time high-resolution practical applications.

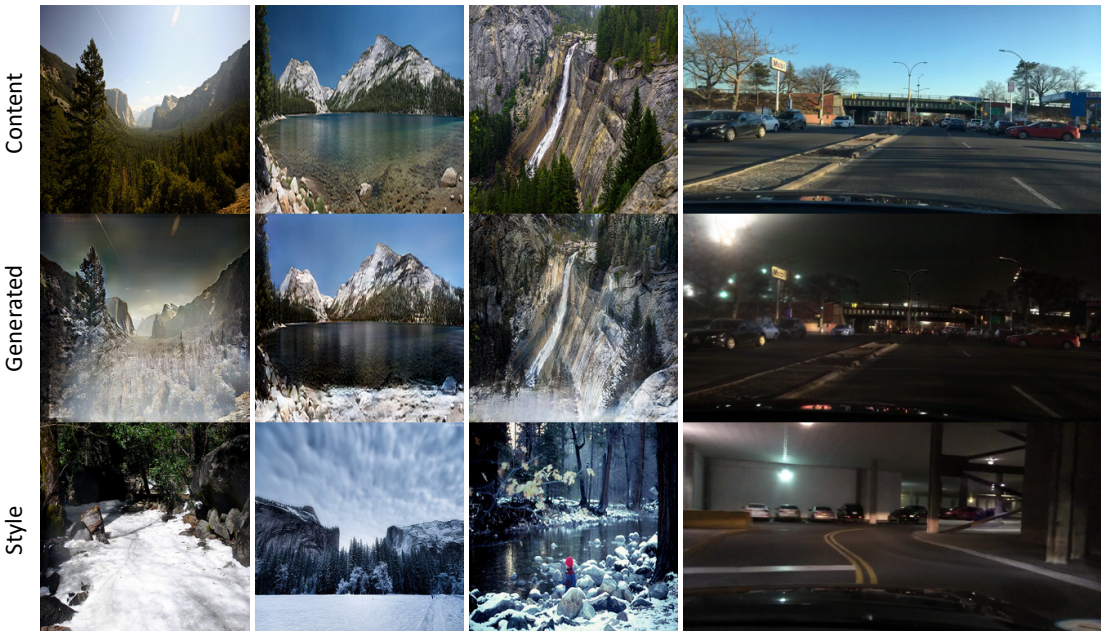

FIGURE 4.11: **Failure cases** (Row 2) generated by the proposed TSIT framework. Some observable artifacts, *e.g.*, unnatural regional stylization and spot artifacts, may exist in arbitrary style transfer.

4.5 Conclusion

We propose TSIT, a simple and versatile framework for image-to-image translation. The proposed symmetrical two-stream network allows the image generation to be effectively conditioned on the multi-scale feature-level semantic structure information and style representation via feature transformations. A systematic study verifies the effectiveness of our method in diverse tasks compared to state-of-the-art task-specific baselines. We believe that designing a unified and versatile framework for more tasks is an important direction in the image generation area. Incorporating unconditional image synthesis tasks and introducing more variability into the two streams/latent space can be interesting future works.

Chapter 5

Focal Frequency Loss for Image Reconstruction and Synthesis

5.1 Introduction

Apart from the progress in the practical perspective of image and video generation (Chapter 3 and Chapter 4), we further wish to tackle the remaining issues through a more fundamental and theoretical study. Thus, we then delve into image generation from a different prospect, *i.e.*, frequency domain. We have seen remarkable progress in image reconstruction and synthesis along with the development of generative models [1, 2, 10, 26, 46], and the progress continues with the emergence of various powerful deep learning-based approaches [4, 34, 168, 169]. Despite their immense success, one could still observe the gaps between the real and generated images in certain cases.

These gaps are sometimes manifested in the form of artifacts that are discernible. For instance, upsampling layers using transposed convolutions tend to produce checkerboard artifacts [170]. The gaps, in some other cases, may only be revealed through the frequency spectrum analysis. Recent studies [128–130] in media forensics have shown some notable periodic patterns in the frequency spectra of manipulated images, which may be consistent with artifacts in the spatial domain. In Figure 5.1, we show some paired examples of real images and the fake ones generated by typical generative models for image reconstruction and synthesis. It is

* The work in this chapter has been published in [60].

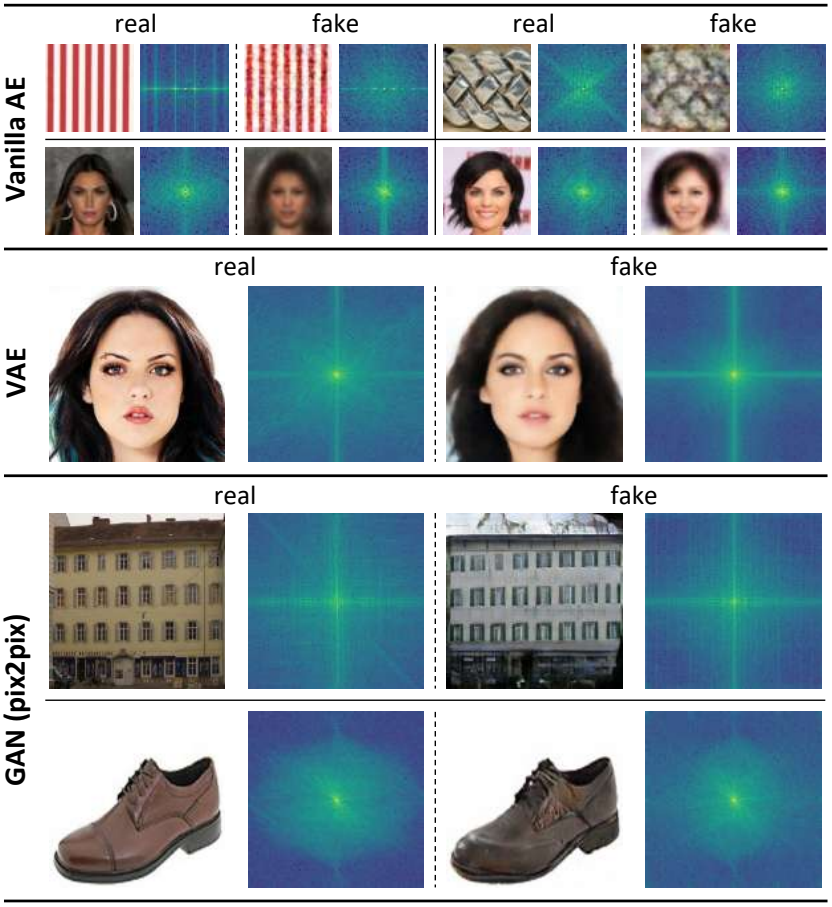

FIGURE 5.1: Frequency domain gaps between the real and the generated images by typical generative models in image reconstruction and synthesis. Vanilla AE [1] loses important frequencies, leading to blurry images (Row 1 and 2). VAE [2] biases to a limited spectrum region (Row 3), losing high-frequency information (outer regions and corners). Unnatural periodic patterns can be spotted on the spectra of images generated by GAN (pix2pix) [3] (Row 4), consistent with the observable checkerboard artifacts (zoom in for view). In some cases, a frequency spectrum region shift occurs to GAN-generated images (Row 5).

observed that the frequency domain gap between the real and fake images may be a common issue for these methods, albeit in slightly different forms.

The observed gaps in the frequency domain may be imputed to some inherent bias of neural networks when applied to reconstruction and synthesis tasks. Fourier analysis highlights a phenomenon called *spectral bias* [37–39], a learning bias of neural networks towards low-frequency functions. Thus, generative models tend to eschew frequency components that are hard to synthesize, *i.e.*, hard frequencies, and converge to an inferior point. *F-Principle* [118] shows that the priority of fitting certain frequencies in a network is also different throughout the training,

usually from low to high. Consequently, it is difficult for a model to maintain important frequency information as it tends to generate frequencies with a higher priority.

In this work, we carefully study the frequency domain gap between real and fake images and explore ways to ameliorate reconstruction and synthesis quality by narrowing this gap. Existing methods [2, 3, 34] usually adopt pixel losses in the spatial domain, while spatial domain losses hardly help a network find hard frequencies and synthesize them, in that every pixel shares the same significance for a certain frequency. In comparison, we transform both the real and generated samples to their frequency representations using the standard discrete Fourier transform (DFT). The images are decomposed into sines and cosines, exhibiting periodic properties. Each coordinate value on the frequency spectrum depends on all the image pixels in the spatial domain, representing a specific spatial frequency. Explicitly minimizing the distance of coordinate values on the real and fake spectra can help networks easily locate difficult regions on the spectrum, *i.e.*, hard frequencies.

To tackle these hard frequencies, inspired by hard example mining [137, 138] and focal loss [139], we propose a simple yet effective frequency-level objective function, named *focal frequency loss*. We map each spectrum coordinate value to a Euclidean vector in a two-dimensional space, with both the amplitude and phase information of the spatial frequency put under consideration. The proposed loss function is defined by the scaled Euclidean distance of these vectors by down-weighting easy frequencies using a dynamic spectrum weight matrix. Intuitively, the matrix is updated on the fly according to a non-uniform distribution on the current loss of each frequency during training. The model will then rapidly focus on hard frequencies and progressively refine the generated frequencies to improve image quality.

The main **contribution** of this work is a novel focal frequency loss that directly optimizes generative models in the frequency domain. We carefully motivate how a loss can be built on a space where frequencies of an image can be well represented and distinguished, facilitating optimization in the frequency dimension. We further explain the way that enables a model to focus on hard frequencies, which may be pivotal for quality improvement. Extensive experiments demonstrate the effectiveness of the proposed loss on representative baselines [1–3, 34], and the loss is complementary to existing spatial domain losses such as perceptual loss [66].

We further show the potential of focal frequency loss to improve state-of-the-art StyleGAN2 [4].

5.2 Focal Frequency Loss

To formulate our method, we explicitly exploit the frequency representation of images (Section 5.2.1), facilitating the network to locate the hard frequencies. We then define a frequency distance (Section 5.2.2) to quantify the differences between images in the frequency domain. Finally, we adopt a dynamic spectrum weighting scheme (Section 5.2.3) that allows the model to focus on the on-the-fly hard frequencies.

5.2.1 Frequency Representation of Images

In this section, we revisit and highlight several key concepts of the discrete Fourier transform. We demonstrate the effect of missing frequencies in the image and the advantage of frequency representation for locating hard frequencies.

Discrete Fourier transform (DFT) is a complex-valued function that converts a discrete finite signal into its constituent frequencies, *i.e.*, complex exponential waves. An image[1] can be treated as a two-dimensional discrete finite signal with only real numbers. Thus, to convert an image into its frequency representation, we perform the 2D discrete Fourier transform:

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot e^{-i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}, \tag{5.1}$$

where the image size is $M \times N$; (x, y) denotes the coordinate of an image pixel in the spatial domain; $f(x, y)$ is the pixel value; (u, v) represents the coordinate of a spatial frequency on the frequency spectrum; $F(u, v)$ is the complex frequency value; e and i are Euler's number and the imaginary unit, respectively. Following Euler's formula:

$$e^{i\theta} = \cos\theta + i\sin\theta, \tag{5.2}$$

[1] For simplicity, the formulas in this section are applied to gray-scale images, while the extension to color images is straightforward by processing each channel separately in the same way.

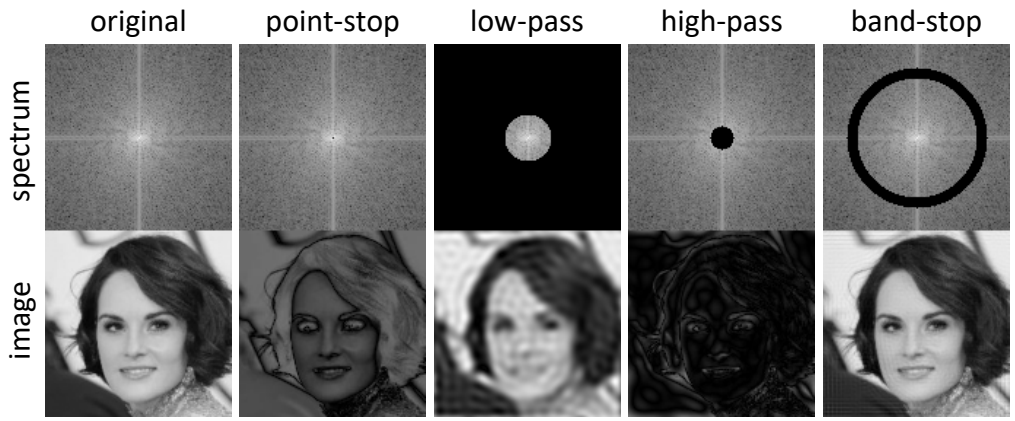

FIGURE 5.2: Standard bandlimiting operations on the frequency spectrum with the origin (low frequencies) center shifted and respective images in the spatial domain. These manual operations can be regarded as a simulation to show the effect of missing frequencies.

the natural exponential function in Eq. (5.1) can be written as:

$$e^{-i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} = \cos 2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right) - i\sin 2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right). \tag{5.3}$$

According to Eq. (5.1) and Eq. (5.3), the image is decomposed into orthogonal sine and cosine functions, constituting the imaginary and the real part of the frequency value, respectively, after applied 2D DFT. Each sine or cosine can be regarded as a binary function of (x, y), where its angular frequency is decided by the spectrum position (u, v). The mixture of these sines and cosines provides both the horizontal and vertical frequencies of an image. Therefore, spatial frequency manifests as the 2D sinusoidal components in the image. The spectrum coordinate (u, v) also represents the angled direction of a spatial frequency, and $F(u, v)$ shows the "response" of the image to this frequency. Due to the periodicity of trigonometric functions, the frequency representation of an image also acquires periodic properties.

Note that in Eq. (5.1), $F(u, v)$ is the sum of a function that traverses every image pixel in the spatial domain, hence a specific spatial frequency on the spectrum depends on all the image pixels. For an intuitive visualization, we suppress the *single* center point (the lowest frequency) of the spectrum (Column 2 of Figure 5.2), leading to *all* the image pixels being affected. To further ascertain the spatial frequency at the different regions on the spectrum, we perform some other standard bandlimiting operations and visualize their physical meanings in the spatial domain (Figure 5.2). A low-pass filter (Column 3), *i.e.*, missing high frequencies, causes

FIGURE 5.3: The necessity of both amplitude and phase information for a frequency distance verified by *single-image reconstruction*.

blur and typical ringing artifacts. A high-pass filter (Column 4), *i.e.*, missing low frequencies, tends to retain edges and boundaries. Interestingly, a simple band-stop filter (Column 5), *i.e.*, missing certain frequencies, produces visible common checkerboard artifacts (zoom in for view).

Observably, the losses of different regions on the frequency spectrum correspond to different artifacts on the image. One may deduce that compensating for these losses may reduce artifacts and improve image reconstruction and synthesis quality. The analysis here shows the value of using the frequency representation of images for profiling and locating different frequencies, especially the hard ones.

5.2.2 Frequency Distance

To devise a loss function for the missing frequencies, we need a distance metric that quantifies the differences between real and fake images in the frequency domain. The distance has to be differentiable to support stochastic gradient descent. In the frequency domain, the data objects are different spatial frequencies on the frequency spectrum, appearing as different 2D sinusoidal components in an image. To design our frequency distance, we further study the real and imaginary part of the complex value $F(u, v)$ in Eq. (5.1).

Let $R(u, v) = a$ and $I(u, v) = b$ be the real and the imaginary part of $F(u, v)$, respectively. $F(u, v)$ can be rewritten as:

$$F(u, v) = R(u, v) + I(u, v)\, i = a + bi. \tag{5.4}$$

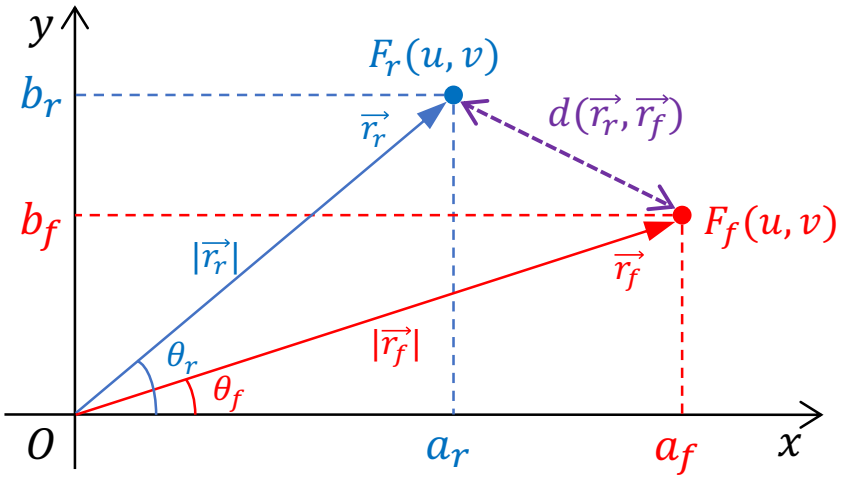

FIGURE 5.4: Frequency distance between $\vec{r_r}$ and $\vec{r_f}$ mapped from two corresponding real and fake frequency values $F_r(u,v)$ and $F_f(u,v)$ at the spectrum position (u,v). The Euclidean distance (purple line) is used, considering both the amplitude (magnitude $|\vec{r_r}|$ and $|\vec{r_f}|$) and phase (angle θ_r and θ_f) information.

According to the definition of 2D discrete Fourier transform, there are two key elements in $F(u,v)$.The first element is *amplitude*, which is defined as:

$$|F(u,v)| = \sqrt{R(u,v)^2 + I(u,v)^2} = \sqrt{a^2+b^2}. \tag{5.5}$$

Amplitude manifests the energy, *i.e.*, how strongly an image responds to the 2D sinusoidal wave with a specific frequency. We typically show the amplitude as an informative visualization of the frequency spectrum (*e.g.*, Figure 5.1 and 5.2). The second element is *phase*, which is written as:

$$\angle F(u,v) = \arctan\left(\frac{I(u,v)}{R(u,v)}\right) = \arctan\frac{b}{a}. \tag{5.6}$$

Phase represents the shift of a 2D sinusoidal wave from the wave with the origin value (the beginning of a cycle).

A frequency distance should consider both the amplitude and the phase as they capture different information of an image. We show a single-image reconstruction experiment in Figure 5.3. Merely minimizing the amplitude difference returns a reconstructed image with irregular color patterns. Conversely, using only the phase information, the synthesized image resembles a noise. A faithful reconstruction can only be achieved by considering both amplitude and phase.

Our solution is to map each frequency value to a Euclidean vector in a two-dimensional space (*i.e.*, a plane). Following the standard definition of a complex number, the real and imaginary parts correspond to the x-axis and y-axis, respectively. Let $F_r(u,v) = a_r + b_r i$ be the spatial frequency value at the spectrum coordinate (u,v) of the real image, and the corresponding $F_f(u,v) = a_f + b_f i$ with the similar meaning *w.r.t.* the fake image. We denote $\vec{r_r}$ and $\vec{r_f}$ as two respective vectors mapped from $F_r(u,v)$ and $F_f(u,v)$ (see Figure 5.4). Based on the definition of amplitude and phase, we note that the vector magnitude $|\vec{r_r}|$ and $|\vec{r_f}|$ correspond to the amplitude, and the angle θ_r and θ_f correspond to the phase. Thus, the frequency distance corresponds to the distance between $\vec{r_r}$ and $\vec{r_f}$, which considers both the vector magnitude and angle. We use the (squared) Euclidean distance for a single frequency:

$$d(\vec{r_r}, \vec{r_f}) = \|\vec{r_r} - \vec{r_f}\|_2^2 = |F_r(u,v) - F_f(u,v)|^2. \tag{5.7}$$

The frequency distance between the real and fake images can be written as the average value:

$$d(F_r, F_f) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} |F_r(u,v) - F_f(u,v)|^2. \tag{5.8}$$

5.2.3 Dynamic Spectrum Weighting

The frequency distance we defined in Eq. (5.8) quantitatively compares the real and fake images in the frequency domain. However, a direct use of Eq. (5.8) as a loss function is not helpful in coping with hard frequencies since the weight of each frequency is identical. A model would still bias to easy frequencies due to the inherent bias.

Inspired by hard example mining [137, 138] and focal loss [139], we formulate our method to focus the training on the hard frequencies. To implement this, we introduce a spectrum weight matrix to down-weight the easy frequencies. The spectrum weight matrix is dynamically determined by a non-uniform distribution on the current loss of each frequency during training. The shape of the matrix is the same as that of the spectrum. The matrix element $w(u,v)$, *i.e.*, the weight for

the spatial frequency at (u, v) is defined as:

$$w\left(u, v\right) = \left|F_r\left(u, v\right) - F_f\left(u, v\right)\right|^{\alpha}, \tag{5.9}$$

where α is the scaling factor for flexibility ($\alpha = 1$ in our experiments). We further normalize the matrix values into the range $[0, 1]$, where the weight 1 corresponds to the currently most lost frequency, and the easy frequencies are down-weighted. The gradient through the spectrum weight matrix is locked, so it only serves as the weight for each frequency.

By performing the Hadamard product for the spectrum weight matrix and the frequency distance matrix, we have the *full form* of the focal frequency loss (FFL):

$$\text{FFL} = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} w\left(u, v\right) \left|F_r\left(u, v\right) - F_f\left(u, v\right)\right|^2. \tag{5.10}$$

The focal frequency loss can be seen as a weighted average of the frequency distance between the real and fake images. It focuses the model on synthesizing hard frequencies by down-weighting easy frequencies. Besides, the focused region is updated on the fly to complement the immediate hard frequencies, thus progressively refining the generated images and being adaptable to different methods.

In practice, to apply the proposed focal frequency loss to a model, we first transform both the real and fake images into their frequency presentations using the 2D DFT. We then perform the orthonormalization for each frequency value $F\left(u, v\right)$, *i.e.*, dividing it by $\sqrt{MN}$, so that the 2D DFT is unitary to ensure a smooth gradient. Finally, we employ Eq. (5.10) to calculate the focal frequency loss.

5.3 Experiments

5.3.1 Settings

Baselines. We start from image reconstruction by vanilla AE [1] (*i.e.*, a simple 2-layer MLP) and VAE [2] (*i.e.*, CNN-based). We then study unconditional image synthesis using VAE, *i.e.*, generating images from the Gaussian noise. Besides, we

also investigate conditional image synthesis using GAN-based methods. Specifically, we select two typical image-to-image translation approaches, *i.e.*, pix2pix [3] and SPADE [34]. We further explore the potential of focal frequency loss (FFL) on state-of-the-art StyleGAN2 [4]. In addition, we compare FFL with relevant losses [66, 67].

Datasets. We use a total of seven datasets. The datasets vary in types, sizes, and resolutions. For vanilla AE, we exploit the Describable Textures Dataset (DTD) [171] and CelebA [172]. For VAE, we use CelebA and CelebA-HQ [49] with different resolutions. For pix2pix, we utilize the officially prepared CMP Facades [173] and edges → shoes [174] datasets. For SPADE, we select two challenging datasets, *i.e.*, Cityscapes [163] and ADE20K [164]. For StyleGAN2, we reuse CelebA-HQ.

Evaluation metrics. To evaluate frequency domain difference, we introduce a frequency-level metric, named Log Frequency Distance (LFD), which is defined by a modified version of Eq. (5.8):

$$\text{LFD} = \log \left[\frac{1}{MN} \left(\sum_{u=0}^{M-1} \sum_{v=0}^{N-1} |F_r(u,v) - F_f(u,v)|^2 \right) + 1 \right], \tag{5.11}$$

where the logarithm is only used to scale the value into a reasonable range. A lower LFD is better. Note that LFD is a full reference metric (*i.e.*, requiring the ground truth image), so we use it in the reconstruction tasks.

Besides, we integrate the evaluation protocols from prior works [13, 21, 34, 38]. Specifically, we employ FID (lower is better) [8] for all tasks. For the reconstruction tasks of vanilla AE and VAE, we use PSNR (higher is better), SSIM (higher is better) [175], and LPIPS (lower is better) [176] in addition to LFD and FID. For the synthesis tasks of VAE, pix2pix, and StyleGAN2, we apply IS (higher is better) [51] in addition to FID. For SPADE (task-specific method for semantic image synthesis), besides FID, we follow their paper [34] to use mIoU (higher is better) and pixel accuracy (accu, higher is better) for the segmentation performance of synthesized images. We use DRN-D-105 [165] for Cityscapes and UperNet101 [166] for ADE20K.

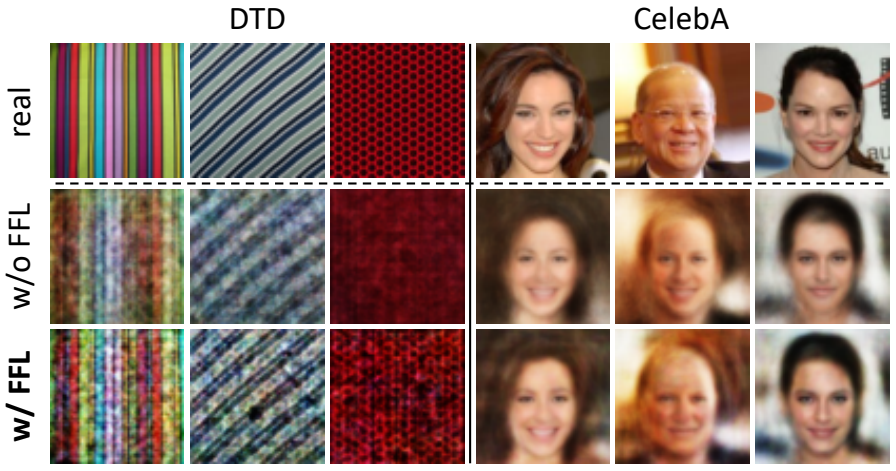

FIGURE 5.5: **Vanilla AE image reconstruction** results on the **DTD** (64×64) and **CelebA** (64 × 64) datasets.

TABLE 5.1: The PSNR (higher is better), SSIM (higher is better), LPIPS (lower is better), FID (lower is better) and LFD (lower is better) scores for the **vanilla AE image reconstruction** trained with/without the focal frequency loss (FFL).

Dataset	FFL	PSNR↑	SSIM↑	LPIPS↓	FID↓	LFD↓
DTD	w/o	20.133	**0.407**	0.414	246.870	14.764
	w/	**20.151**	0.400	**0.404**	**240.373**	**14.760**
CelebA	w/o	20.044	0.568	0.237	97.035	14.785
	w/	**21.703**	**0.642**	**0.199**	**83.801**	**14.403**

5.3.2 Results and Analysis

Vanilla AE. The results of vanilla AE [1] image reconstruction are shown in Figure 5.5. On DTD, without the focal frequency loss (FFL), the vanilla AE baseline synthesizes blurry images, which lack sufficient texture details and only contain some low-frequency information. With FFL, the reconstructed images become clearer and show more texture details. The results on CelebA show that FFL improves a series of quality problems, *e.g.*, face blur (Column 4), identity shift (Column 5), and expression loss (Column 6).

The quantitative evaluation results are presented in Table 5.1. Adding the proposed FFL to the vanilla AE baseline leads to a performance boost in most cases on the DTD and CelebA datasets *w.r.t.* five evaluation metrics. We note that the performance boost on CelebA is larger, indicating the effectiveness of FFL for the natural images.

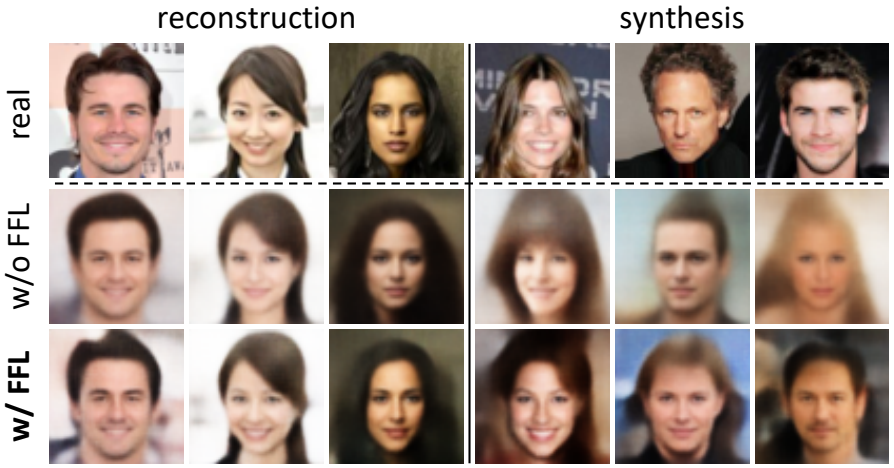

FIGURE 5.6: **VAE image reconstruction** and **unconditional image synthesis** results on the **CelebA** (64×64) dataset.

TABLE 5.2: The PSNR (higher is better), SSIM (higher is better), LPIPS (lower is better), FID (lower is better) and LFD (lower is better) scores for the **VAE image reconstruction** trained with/without the focal frequency loss (FFL).

Dataset	FFL	PSNR↑	SSIM↑	LPIPS↓	FID↓	LFD↓
CelebA	w/o	19.961	0.606	0.217	69.900	14.804
	w/	**22.954**	**0.723**	**0.143**	**49.689**	**14.115**
CelebA-HQ	w/o	21.310	0.616	0.367	71.081	17.266
	w/	**22.253**	**0.637**	**0.344**	**59.470**	**17.049**

VAE. The results of VAE [2] image reconstruction and unconditional image synthesis on CelebA are shown in Figure 5.6. For reconstruction, FFL helps the VAE model better retain the image clarity (Column 1), expression (Column 2), and skin color (Column 3). The unconditional synthesis results (Column 4, 5, 6) show that the quality of generated images is improved after applying FFL. The generated faces become clearer and gain more texture details. For a higher resolution, we present the VAE reconstruction and synthesis results on CelebA-HQ in Figure 5.7. By adding FFL to the VAE baseline, the reconstructed images keep more original image information, *e.g.*, mouth color (Column 2) and opening angle (Column 1). Besides, high-frequency details on the hair are clearly enhanced (Column 1). For unconditional image synthesis, FFL helps reduce artifacts and ameliorates the perceptual quality of synthesized images.

The quantitative test results of VAE image reconstruction are shown in Table 5.2. Adding FFL to the VAE baseline achieves better performance *w.r.t.* all the metrics. Besides, both FID and IS are better in the unconditional image synthesis task

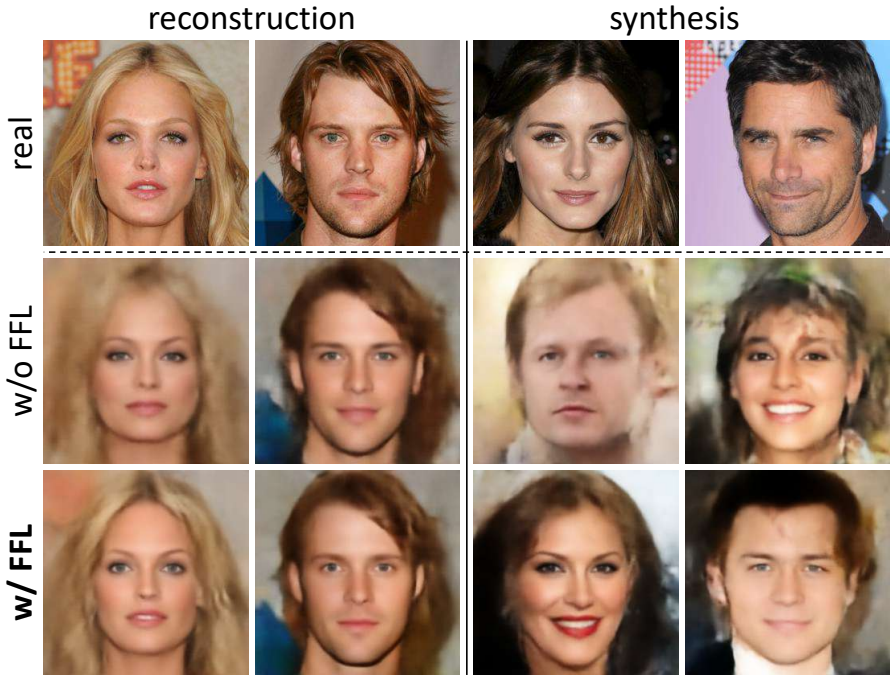

FIGURE 5.7: **VAE image reconstruction** and **unconditional image synthesis** results on the **CelebA-HQ** (256×256) dataset.

TABLE 5.3: The FID (lower is better) and IS (higher is better) scores for the **VAE unconditional image synthesis** trained with/without the focal frequency loss (FFL).

Dataset	FFL	FID↓	IS↑
CelebA	w/o	80.116	1.873
	w/	**71.050**	**2.010**
CelebA-HQ	w/o	93.778	2.057
	w/	**84.472**	**2.060**

(Table 5.3), indicating that the generated images are clearer and more photorealistic. The results suggests the effectiveness of the focal frequency loss in helping VAE to improve image reconstruction and synthesis quality.

pix2pix. For conditional image synthesis, the results of pix2pix [3] image-to-image translation (I2I) are shown in Figure 5.8. On CMP Facades, FFL improves the image synthesis quality of pix2pix by reducing unnatural colors (Column 1) or the black artifacts on the building (Column 2). Meanwhile, the semantic information alignment with the mask becomes better after applying FFL. For the edges → shoes translation, pix2pix baseline sometimes introduces colored checkerboard artifacts to the white background (Column 3, zoom in for view). Besides, atypical colors appear in certain cases (Column 4). In comparison, the model trained with FFL

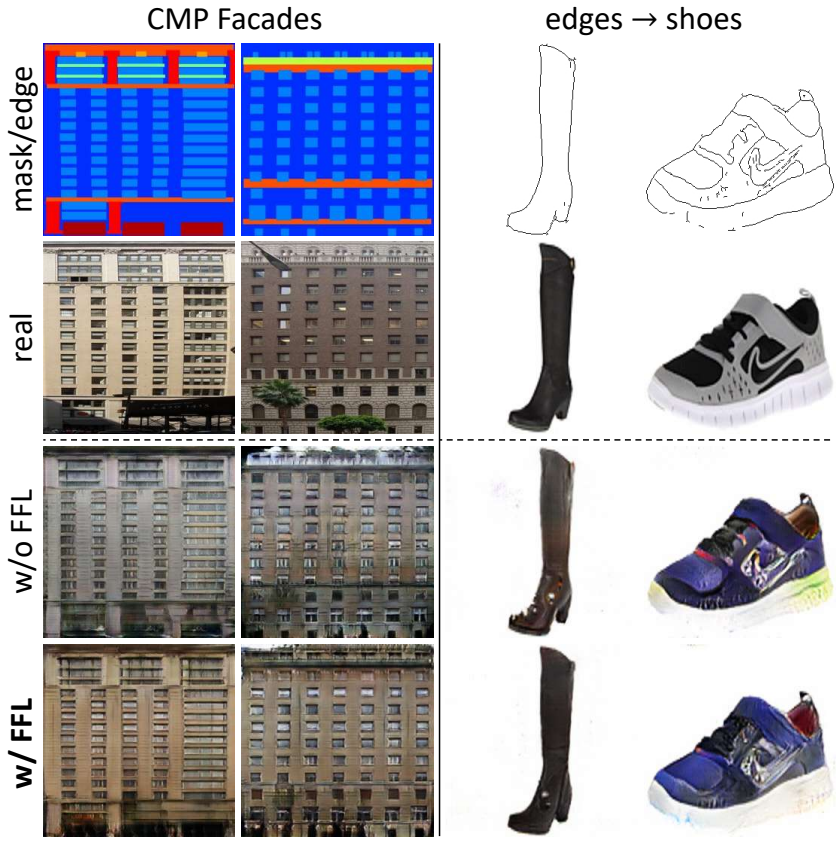

FIGURE 5.8: **pix2pix image-to-image translation** results on **CMP Facades** (256 × 256) and **edges → shoes** (256 × 256) datasets.

TABLE 5.4: The FID (lower is better) and IS (higher is better) scores for the **pix2pix image-to-image translation** trained with/without the focal frequency loss (FFL).

Dataset	FFL	FID↓	IS↑
CMP Facades	w/o	128.492	1.571
	w/	**123.773**	**1.738**
edges → shoes	w/o	80.279	2.674
	w/	**74.359**	**2.804**

yields fewer artifacts.

The quantitative evaluation results of pix2pix image-to-image translation are shown in Table 5.4. FFL contributes to a performance boost on both of the two datasets. The results of the pix2pix baseline show the adaptability of the focal frequency loss for the image-to-image translation problem.

SPADE. We further explore semantic image synthesis (*i.e.*, synthesizing a photo-realistic image from a semantic segmentation mask) on more challenging datasets. The results of SPADE [34] are shown in Figure 5.9. In the street scene of Cityscapes

FIGURE 5.9: **SPADE semantic image synthesis** results on the **Cityscapes** (512 × 256) and **ADE20K** (256 × 256) datasets.

TABLE 5.5: The mIoU (higher is better), pixel accuracy (accu, higher is better) and FID (lower is better) scores for the **SPADE semantic image synthesis** trained with/without the focal frequency loss (FFL) compared to a series of task-specific methods.

	Cityscapes			ADE20K		
Method	mIoU↑	accu↑	FID↓	mIoU↑	accu↑	FID↓
CRN [32]	52.4	77.1	104.7	22.4	68.8	73.3
SIMS [65]	47.2	75.5	**49.7**	N/A	N/A	N/A
pix2pixHD [33]	58.3	81.4	95.0	20.3	69.2	81.8
SPADE [34]	62.3	81.9	71.8	38.5	79.9	33.9
SPADE + FFL	**64.2**	**82.5**	59.5	**42.9**	**82.4**	**33.7**

(Column 1), SPADE baseline distorts the car and road, missing some important details (*e.g.*, road line). The model trained with FFL demonstrates better perceptual quality for these details. In the outdoor scene of ADE20K (Column 2), applying FFL to SPADE boosts its ability to generate details on the buildings. Besides, for the ADE20K indoor images (Column 3), SPADE baseline produces some abnormal artifacts in certain cases. The model trained with the proposed FFL synthesizes more photorealistic images.

FIGURE 5.10: **StyleGAN2 unconditional image synthesis** results (without truncation) and the mini-batch average spectra (adjusted to better contrast) on the **CelebA-HQ** (256×256) dataset.

TABLE 5.6: The FID (lower is better) and IS (higher is better) scores for the **StyleGAN2 unconditional image synthesis** trained with/without the focal frequency loss (FFL).

Dataset	FFL	FID↓	IS↑
CelebA-HQ	w/o	5.696	3.383
(256×256)	w/	**4.972**	**3.432**

The quantitative test results are presented in Table 5.5 (the values used for comparison are taken from [34]). We compare SPADE trained with/without FFL against a series of open-source task-specific semantic image synthesis methods [32, 33, 65]. SIMS [65] obtains the best FID but poor segmentation scores on Cityscapes in that it directly stitches the training image patches from a memory bank while not keeping the exactly consistent positions. Without modifying the SPADE network structure, training with FFL contributes a further performance boost, greatly outperforming the benchmark methods, which suggests the effectiveness of FFL for semantic image synthesis.

StyleGAN2. We apply FFL to the mini-batch average spectra of the real images and the generated images by the state-of-the-art unconditional image synthesis method, *i.e.*, StyleGAN2 [4], intending to narrow the frequency distribution gap and improve quality further. The results on CelebA-HQ (256×256) without truncation [4, 5] are shown in Figure 5.10. Although StyleGAN2 (w/o FFL) generates photorealistic images in most cases, some tiny artifacts can still be spotted on the background (Column 2 and 4) and face (Column 5). Applying FFL, such

TABLE 5.7: **Comparison** of our focal frequency loss (FFL) **with relevant losses**, *i.e.*, perceptual loss (PL), spectral regularization (SpReg), and another transformation form for FFL, *i.e.*, discrete cosine transform (DCT), in different image reconstruction and synthesis tasks. (a) VAE image reconstruction (CelebA). (b) VAE unconditional image synthesis (CelebA). (c) pix2pix image-to-image translation (edges → shoes).

	(a)					(b)		(c)	
Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	LFD↓	FID↓	IS↑	FID↓	IS↑
baseline	19.961	0.606	0.217	69.900	14.804	80.116	1.873	80.279	2.674
+ PL [66]	20.964	0.658	**0.143**	62.795	14.573	78.825	1.788	78.916	2.722
+ SpReg [67]	19.974	0.607	0.218	69.118	14.796	78.079	1.898	79.300	2.700
+ FFL (DCT)	22.677	0.711	0.150	51.536	14.179	71.827	1.932	79.045	2.754
+ FFL (Ours)	**22.954**	**0.723**	**0.143**	**49.689**	**14.115**	**71.050**	**2.010**	**74.359**	**2.804**

artifacts are reduced, ameliorating synthesis quality further. Observably, the frequency domain gaps between mini-batch average spectra are clearly mitigated by FFL (Column 8).

The quantitative results are reported in Table 5.6. FFL improves both FID and IS, in line with the visual quality enhancement. The results on StyleGAN2 show the potential of FFL to boost state-of-the-art baseline performance.

Comparison with relevant losses. For completeness and fairness, we compare the proposed focal frequency loss (FFL) with relevant loss functions that aim at improving image reconstruction and synthesis quality. Specifically, we select the widely used spatial-based method, *i.e.*, perceptual loss (PL) [66], which depends on high-level features from a pre-trained VGG [161] network. We also study the frequency-based approach, *i.e.*, spectral regularization (SpReg) [67], which is derived based on the azimuthal integration of the Fourier power spectrum. Besides, we further compare with another transformation form for FFL, *i.e.*, discrete cosine transform (DCT).

The comparison results are reported in Table 5.7. FFL outperforms the relevant approaches (*i.e.*, PL and SpReg) when applied to our baselines in different image reconstruction and synthesis tasks. It is noteworthy that FFL and PL are complementary, as shown by our previous experiments on SPADE, which also uses PL. Even if we replace DFT with DCT as the transformation form of FFL, the results are still better than previous methods. The performance is only slightly inferior to that obtained by FFL with DFT (*i.e.*, Eq. (5.10)). We deduce that the

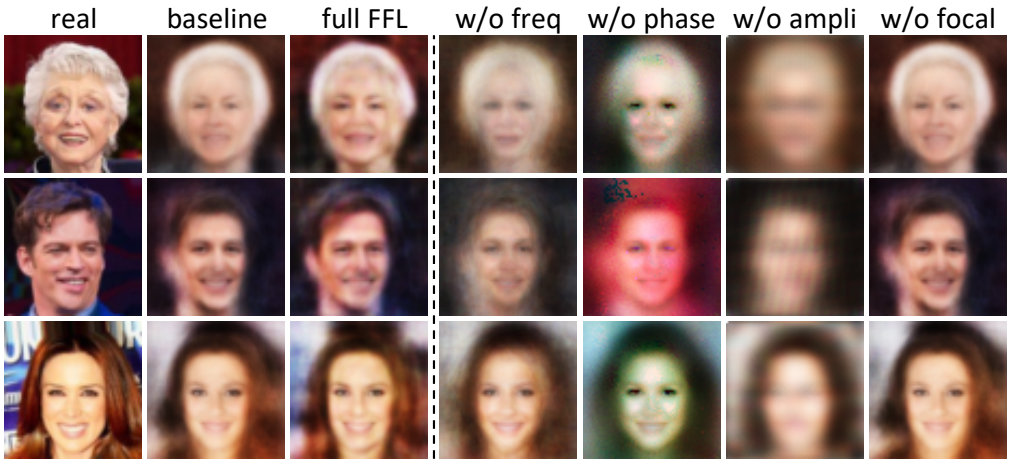

FIGURE 5.11: **Ablation studies** of each key component for the focal frequency loss (FFL), *i.e.*, frequency representation (freq), phase and amplitude (ampli) information, and dynamic spectrum weighting (focal) in the vanilla AE image reconstruction task on CelebA.

TABLE 5.8: The PSNR (higher is better), SSIM (higher is better), LPIPS (lower is better), FID (lower is better) and LFD (lower is better) scores for the **ablation studies** of each key component for the focal frequency loss (FFL).

	PSNR↑	SSIM↑	LPIPS↓	FID↓	LFD↓
baseline	20.044	0.568	0.237	97.035	14.785
full FFL	**21.703**	**0.642**	**0.199**	**83.801**	**14.403**
w/o freq	18.200	0.470	0.265	123.833	15.210
w/o phase	13.273	0.380	0.407	233.170	16.344
w/o ampli	15.640	0.359	0.539	323.528	15.799
w/o focal	20.163	0.574	0.234	94.497	14.758

transformation form for FFL may be flexible. At this stage, DFT may be a better choice.

Ablation studies. We present ablation studies of each key component for FFL in Figure 5.11 and corresponding quantitative results in Table 5.8. For intuitiveness and simplicity, we use vanilla AE image reconstruction on CelebA for the evaluation.

The full FFL achieves the best performance. If we do not use the frequency representation of images (Section 5.2.1) and focus the model on hard pixels in the spatial domain, the synthesized images become more blurry. The quantitative results degrade. Discarding either the phase or amplitude information (Section 5.2.2) harms the metric performance vastly. Visually, using no phase information (amplitude only), the contour of reconstructed faces is retained, but the color is completely shifted. Without amplitude (phase only), the model cannot reconstruct the faces at all, and the full identity information is lost. This further verifies the necessity

of considering both amplitude and phase information. Without focusing the model on the hard frequencies by the dynamic spectrum weighting (*i.e.*, directly using Eq. (5.8)), the results are visually similar to baseline, in line with our discussion in Section 5.2.3. The metrics decrease, being close to but slightly better than baseline, which may benefit from the frequency representation.

5.4 Conclusion

The proposed focal frequency loss directly optimizes image reconstruction and synthesis methods in the frequency domain. The loss adaptively focuses the model on the frequency components that are hard to deal with to ameliorate quality. The loss is complementary to existing spatial losses of diverse baselines varying in categories, network structures, and tasks, outperforming relevant approaches. We further show the potential of focal frequency loss to improve the synthesis results of StyleGAN2. Exploring other applications and devising better frequency domain optimization strategies can be interesting future works. For instance, aside from devising a frequency-level loss function, exploring frequency representations for the network inputs could be interesting. For GAN-based methods, we may also consider a frequency discriminator as an auxiliary measure to the spatial one. For future applications, we may further explore the potential of focal frequency loss on state-of-the-art diffusion models and additional downstream tasks.

Chapter 6

Deceive D: Adaptive Pseudo Augmentation for GAN Training with Limited Data

6.1 Introduction

Thanks to our exploration of the aforementioned useful applications (Chapter 3 and Chapter 4) and fundamental study (Chapter 5) in the field of image and video generation, the performance of various generative models can be evidently improved, in terms of both the fidelity and diversity of synthesized data. However, current generative models like GANs [4, 10] usually require a large amount of training data to fully unleash their power. Training GANs with insufficient data tends to generate poor-quality images, as shown in Figure 6.1. Collecting sufficient data samples for these GANs is sometimes infeasible, especially in domains where data are sparse and privacy-sensitive. To ease the practical deployment of powerful GANs, it is necessary to devise new strategies for training GANs with limited data while preserving the quality of synthesis.

Recent studies have shown that the overfitting of the discriminator is the critical reason that impedes effective GAN training on limited data [40–43], rendering severe instability of training dynamics. Specifically, when the discriminator starts to

* The work in this chapter has been published in [68].

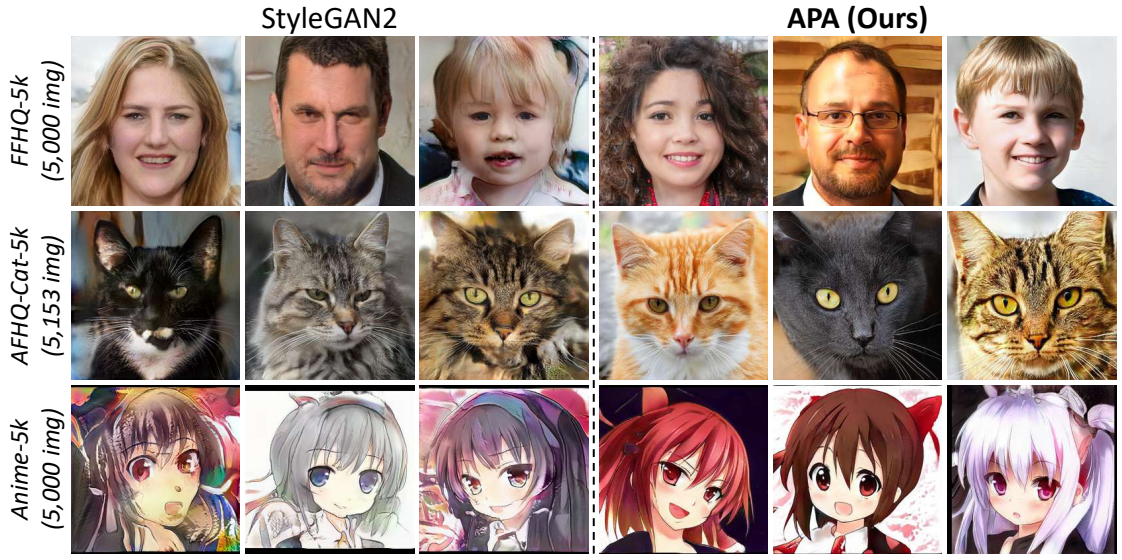

FIGURE 6.1: **StyleGAN2** [4] synthesized results (no truncation) deteriorate **given the limited amount of training data** (256×256), *i.e.*, FFHQ [5] (a subset of $5,000$ images, $\sim 7\%$ of full data), AFHQ-Cat [6] ($5,153$ images, which is small by itself), and Danbooru2019 Portraits (Anime) [7] (a subset of $5,000$ images, $\sim 2\%$ of full data). The proposed Adaptive Pseudo Augmentation (APA) effectively ameliorates the degraded performance of StyleGAN2 on limited data.

overfit, the distributions of its outputs for real and generated samples gradually diverge from each other [41, 43], and its feedback to the generator becomes less informative. Consequently, the generator converges to an inferior point, compromising the quality of synthesized images. Recent solutions to this problem include the use of standard data augmentations, either conventional or differentiable, to real and generated images [41, 42, 44, 45] or applying an additional model regularization term [43]. Addressing the discriminator overfitting is still an open problem. We are interested in finding an alternative way to the aforementioned approaches.

In this work, we present a simple yet effective way to regularize a discriminator without introducing any external augmentations or regularization terms. We call our method *Adaptive Pseudo Augmentation* (APA). In contrast to previous standard data augmentations [41, 42, 44, 45], we exploit the generator in a GAN itself to provide the augmentation, a more natural way to regularize the overfitting of the discriminator. Compared to the model regularization, our approach is more adaptive to fit different settings and training status without manual tuning. Specifically, APA takes the fake/pseudo samples synthesized by the generator and moderately feeds them into the limited real data. Such pseudo data are adaptively presented to the discriminator as "real" instances. The goal of this pseudo augmentation for

the real data is not to enlarge the real dataset but to suppress the discriminator's confidence in distinguishing real and fake distributions. The deceit is introduced adaptively, which is moderated by a deception probability according to the degree of overfitting. To quantify overfitting, we study a series of plausible heuristics derived from the discriminator raw output logits.

The main **contribution** of this work is a novel adaptive pseudo augmentation method for training GANs with limited data. This approach deceives the discriminator adaptively and mitigates the problem of discriminator overfitting. The proposed APA can be readily added to existing GAN training with negligible computational cost. We conduct extensive experiments to demonstrate the effectiveness of APA for state-of-the-art GAN training with limited data. The results are comparable or even better than other types of solutions [41, 43]. APA is also complementary to existing methods based on standard data augmentations for gaining a further performance boost. Besides, we theoretically connect APA with minimizing the JS divergence [54] between the smoothed data distribution and generated distribution, proving its convergence and rationality. We hope that our approach could extend the breadth and potential of solutions to GAN training with limited data.

6.2 Methodology

In GAN's adversarial training, the goal of the generator G is to deceive the discriminator D and maximize the probability that D makes a wrong judgment. Therefore, G keeps refining its generated samples to better deceive D over time. When the training only accesses a limited amount of data, one would observe that D turns out to be overly confident and hardly makes any mistake, causing its feedback to G to become meaningless.

In Figure 6.2, we show the training "snapshots" of two StyleGAN2 [4] models on the FFHQ dataset [5]. The settings of the two models differ only by the amount of data available to them for training. As can be observed, both training processes start smoothly, and distributions of discriminator outputs for the real and generated images overlap at the early stage. As the training progresses, the discriminator, which only has access to limited data (7k images, 10% of full data), experiences diverged

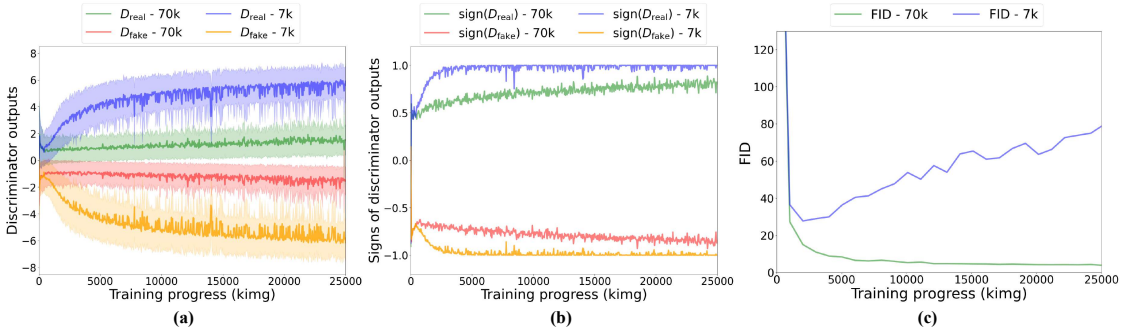

FIGURE 6.2: The overfitting of discriminator in GANs when limited training data are available. The three subplots report statistics of training snapshots of two StyleGAN2 [4] models on FFHQ [5] (256×256). "70k" indicates the full dataset, and "7k" means a subset of $7,000$ images (10% data). The "kimg" denotes thousands of real images shown to the discriminator. (a) Discriminator raw output logits. (b) Signs of discriminator outputs. (c) Training convergence measured by FID [8].

predictions much more rapidly, and the average sign boundary turns out to be more apparent. This divergence in prediction shows that D becomes increasingly confident in classifying real and fake samples. At the late stage of training, D can even judge all the input samples correctly with high confidence. Meanwhile, the evaluation FID [8] scores (lower is better) deteriorate, consistent with the divergence of D's predictions. The phenomena above demonstrate how a discriminator gets overfitted quickly with limited data. As can be seen from the FID curves, the overfitting directly influences training dynamics and the convergence of G, leading to poor generation performance.

6.2.1 Adaptive Pseudo Augmentation

The generator itself naturally possesses the capability to counteract discriminator overfitting. To harness this capability, our method employs a GAN to augment itself using the generated samples (see Figure 6.3). Specifically, we feed the samples generated by G into the limited real data to form a pseudo-real set. The fake images in this set will be adaptively presented to D, pretending themselves as real data. The goal here is to deceive D with the pseudo-real set and consequently suppressing its confidence in distinguishing real and fake images.

Blindly presenting the fake images as reals to D may weaken the fundamental ability of D in adversarial training. In our approach, the deceit is introduced

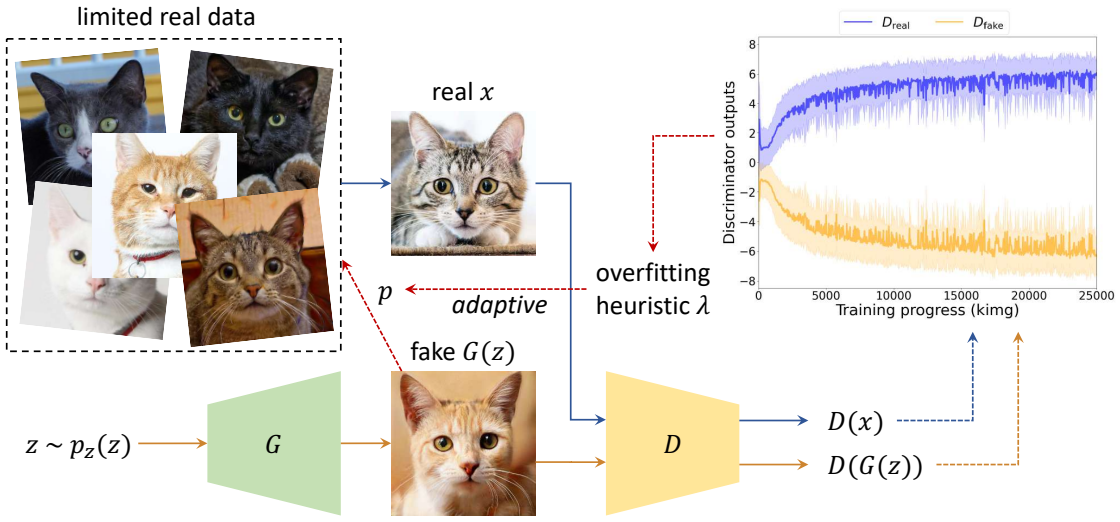

FIGURE 6.3: Adaptive pseudo augmentation (APA) for GAN training with limited data. We employ a GAN to augment itself using the generated images to deceive the discriminator adaptively. Specifically, APA feeds the images synthesized by the generator into the limited real data moderately, and these fakes are presented as "real" instances to the discriminator. Such deceits are introduced adaptively using an overfitting heuristic λ defined by the discriminator raw output logits. The augmentation/deception probability p can be adaptively controlled throughout training.

adaptively to avoid any potential adverse effects. To moderate the deception, we perform pseudo augmentation based on a probability, $p \in [0, 1)$, that quantifies the deception strength. Specifically, the pseudo augmentation will be applied with the probability p or be skipped with the probability $1 - p$.

We note that the overfitting state of D is dynamic throughout the training (see Figure 6.2). It is intuitive to let the deception probability p be adjusted adaptively based on the degree of overfitting. Ideally, p should be adjusted without manual tuning regardless of data scales and properties. To achieve this goal, inspired by ADA [41], we apply an overfitting heuristic λ that quantifies the degree of D's overfitting. We extend the control scheme of ADA and provide three plausible variants:

$$\lambda_r = \mathbb{E}\left(\text{sign}\left(D_{\text{real}}\right)\right), \quad \lambda_f = -\mathbb{E}\left(\text{sign}\left(D_{\text{fake}}\right)\right), \quad \lambda_{rf} = \frac{\mathbb{E}\left(\text{sign}\left(D_{\text{real}}\right)\right) - \mathbb{E}\left(\text{sign}\left(D_{\text{fake}}\right)\right)}{2}, \tag{6.1}$$

where D_{real} and D_{fake} are defined as

$$D_{\text{real}} = \text{logit}\left(D\left(x\right)\right), \quad D_{\text{fake}} = \text{logit}\left(D\left(G\left(z\right)\right)\right), \tag{6.2}$$

where logit denotes the logit function. As shown in Figure 6.2, the λ_r in Eq. (6.1) estimates the portion of real images that obtain positive logit predictions by D, and that of generated images is captured by λ_f. Besides, λ_{rf} indicates half of the distance between the signs of the real and fake logits. For all these heuristics, $\lambda = 0$ represents no overfitting, and $\lambda = 1$ means complete overfitting. We use λ_r in our main experiments and study other variants in the ablation study.

The strategy of using λ to adjust p is as follows. We set a threshold t (in most cases of our experiments, $t = 0.6$) and initialize p to be zero. If λ signifies too much/little overfitting regarding t (*i.e.*, larger/smaller than t), the probability p will be increased/decreased by one fixed step. Using this step size, p can increase from zero to one in 500k images shown to D. We adjust p once every four iterations and clamp p from below to zero after each adjustment. In this way, the strength of pseudo augmentation can be adaptively controlled based on the degree of overfitting (see Figure 6.3).

6.2.2 Theoretical Analysis

Let $p_z(z)$ be the prior on the input noise variable. The mapping from the latent space to the image space is denoted as $G(z)$. For sample x, $D(x)$ represents the estimated probability of x coming from the real data. To examine the rationality of APA, we analyze it in a non-parametric setting [10], where a model is represented with infinite capacity by exploring its convergence in the space of probability density functions. Ideally, the estimated probability distribution p_g defined by G should perfectly model the real data distribution p_{data} without bias if given enough capability and training time.

Since the deception strength p is adaptively adjusted, to facilitate this theoretical analysis, we make a mild assumption that α is the expected strength that approximates the effect of dynamic adjustment during the entire training procedure. Since $p \in [0, 1)$, we have $0 \leq \alpha < p_{\max} < 1$, where $p_{\max}$ is the maximum of deception strength throughout training. Hence, the value function $V(G, D)$ for the minimax

two-player game of APA can be reformulated as:

$$\min_G \max_D V(G,D) = (1-\alpha)\,\mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log D(x)\right] + \alpha\;\mathbb{E}_{z\sim p_z(z)}\left[\log D(G(z))\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1-D(G(z))\right)\right], \tag{6.3}$$

First, let us consider the optimal discriminator [10] for any given generator.

Proposition 1. *If the generator G is fixed, the optimal discriminator D for APA is:*

$$D_G^*(x) = \frac{(1-\alpha)\,p_{\text{data}}(x) + \alpha\;p_g(x)}{(1-\alpha)\,p_{\text{data}}(x) + (1+\alpha)\,p_g(x)} \tag{6.4}$$

Proof. Applying APA, given any generator G, the training objective of the discriminator D is to maximize the value function $V(G,D)$ in Eq. (6.3):

$$\begin{aligned} V(G,D) &= (1-\alpha)\int_x p_{\text{data}}(x)\log D(x)dx + \alpha\int_z p_z(z)\log D(g(z))dz + \int_z p_z(z)\log\left(1-D(g(z))\right)dz \\ &= \int_x \left[(1-\alpha)\,p_{\text{data}}(x)\log D(x) + \alpha\;p_g(x)\log D(x) + p_g(x)\log\left(1-D(x)\right)\right]dx \\ &= \int_x \left[\left((1-\alpha)\,p_{\text{data}}(x) + \alpha\;p_g(x)\right)\log D(x) + p_g(x)\log\left(1-D(x)\right)\right]dx \end{aligned} \tag{6.5}$$

For any $(m,n)\in\mathbb{R}^2\setminus\{0,0\}$, the function $f(y) = m\log(y) + n\log(1-y)$ achieves its maximum in the range $[0,1]$ at $\frac{m}{m+n}$. Besides, the discriminator D is defined only inside of $\text{supp}(p_{\text{data}})\cup\text{supp}(p_g)$, where supp is the set-theoretic support. Therefore, we conclude the proof for Proposition 1.

We have got the optimal discriminator $D_G^*(x)$ in Eq. (6.4) that maximizes the value function $V(G,D)$ given any fixed generator G. The goal of generator G in the adversarial training is to minimize the value function $V(G,D)$ in Eq. (6.3) when D achieves the optimum. Since the training objective of D can be interpreted as maximizing the log-likelihood for the conditional probability $P(Y=y|x)$, where Y estimates that x comes from p_{data} (*i.e.*, $y=1$) or from p_g (*i.e.*, $y=0$), we reformulate virtual training criterion [10] as:

$$\begin{aligned} C(G) &= (1-\alpha)\,\mathbb{E}_{x\sim p_{\text{data}}}\left[\log D_G^*(x)\right] + \alpha\;\mathbb{E}_{z\sim p_z}\left[\log D_G^*(G(z))\right] + \mathbb{E}_{z\sim p_z}\left[\log\left(1-D_G^*(G(z))\right)\right] \\ &= (1-\alpha)\,\mathbb{E}_{x\sim p_{\text{data}}}\left[\log D_G^*(x)\right] + \alpha\;\mathbb{E}_{x\sim p_g}\left[\log D_G^*(x)\right] + \mathbb{E}_{x\sim p_g}\left[\log\left(1-D_G^*(x)\right)\right] \end{aligned} \tag{6.6}$$

Then, let us consider the global minimum of $C(G)$ trained with the proposed APA.

Proposition 2. *Applying APA, the global minimum of the virtual training criterion $C(G)$ is still achieved if and only if $p_g = p_{\text{data}}$, where $C(G) = -\log 4$.*

Proof. 1) If $p_g = p_{\text{data}}$, we have $D_G^*(x) = \frac{1}{2}$ according to Eq. (6.4). By inspecting Eq. (6.6) at $D_G^*(x) = \frac{1}{2}$, we get $C^*(G) = (1-\alpha)\log\frac{1}{2} + \alpha\log\frac{1}{2} + \log\frac{1}{2} = \log\frac{1}{2} + \log\frac{1}{2} = -\log 4$.

2) To verify $C^*(G)$ is the global minimum of $C(G)$, and it can only be achieved when $p_g = p_{\text{data}}$, as in the derivation of Eq. (6.5), we obtain:

$$C(G) = \int_x \left((1-\alpha)\, p_{\text{data}}(x) + \alpha\ p_g(x)\right) \log D_G^*(x) dx + \int_x p_g(x) \log\left(1 - D_G^*(x)\right) dx \tag{6.7}$$

Observe that

$$\begin{aligned} -\log 4 &= (1-\alpha)\, \mathbb{E}_{x\sim p_{\text{data}}}[-\log 2] + \alpha\ \mathbb{E}_{x\sim p_g}[-\log 2] + \mathbb{E}_{x\sim p_g}[-\log 2] \\ &= -\int_x \left((1-\alpha)\, p_{\text{data}}(x) + \alpha\ p_g(x)\right) \log 2\ dx - \int_x p_g(x) \log 2\ dx \end{aligned} \tag{6.8}$$

Subtracting Eq. (6.8) from Eq. (6.7),

$$C(G) = -\log 4 + \int_x \left((1-\alpha)\, p_{\text{data}}(x) + \alpha\ p_g(x)\right) \log 2 \cdot D_G^*(x) dx + \int_x p_g(x) \log 2 \cdot \left(1 - D_G^*(x)\right) dx \tag{6.9}$$

By substituting Eq. (6.4) into Eq. (6.9), we achieve:

$$\begin{aligned} C(G) = -\log 4 &+ \text{KLD}\left(\left((1-\alpha)\, p_{\text{data}} + \alpha\ p_g\right) \middle\| \frac{(1-\alpha)\, p_{\text{data}} + (1+\alpha)\, p_g}{2}\right) \\ &+ \text{KLD}\left(p_g \middle\| \frac{(1-\alpha)\, p_{\text{data}} + (1+\alpha)\, p_g}{2}\right), \end{aligned} \tag{6.10}$$

where KLD is the Kullback-Leibler (KL) divergence. Moreover, Eq. (6.10) further implies that the generation process of G by APA can be regarded as minimizing the Jensen-Shannon (JS) divergence between the smoothed data distribution and the generated distribution:

$$C(G) = -\log 4 + 2 \cdot \text{JSD}\left(\left((1-\alpha)\, p_{\text{data}} + \alpha\ p_g\right) \| p_g\right). \tag{6.11}$$

For the two distributions P and Q, their JS divergence $\text{JSD}(P\|Q) \geq 0$ and $\text{JSD}(P\|Q) = 0$ if and only if $P = Q$. Therefore, for $0 \leq \alpha < p_{\max} < 1$, we obtain that $C^*(G) = -\log 4$ is the global minimum of $C(G)$, and the only solution is $(1-\alpha)\, p_{\text{data}} + \alpha\ p_g = p_g$, *i.e.*, $p_g = p_{\text{data}}$. Q.E.D.

Given the proof in [10], if G and D have enough capacity to reach their optimum, Proposition 2 indicates that the generated distribution p_g can ideally converge to the real data distribution p_{data}. So far, we have proved the convergence of G trained with our proposed APA, which can perfectly model the real data distribution given sufficient capability and training time. Besides, the JS divergence term between the smoothed data distribution and the generated distribution in Eq. (6.11) implies that the judgment of D may be moderated to alleviate overfitting. These conclusions explain the rationality of the proposed APA for training GANs with limited data.

6.3 Experiments

Datasets. We use four datasets in main experiments: Flickr-Faces-HQ (FFHQ) [5] with $70,000$ human face images, AFHQ-Cat [6] with $5,153$ cat faces, Caltech-UCSD Birds-200-2011 (CUB) [9] with $11,788$ images of birds, and Danbooru2019 Portraits (Anime) [7] with $302,652$ anime portrait images. We exploit some of their artificially limited subsets under different settings. All the images are resized to a moderate resolution of 256×256 using a high-quality Lanczos filter [177] to reduce the energy consumption for large-scale GAN training while preserving image quality.

Evaluation metrics. We follow the standard evaluation protocol [13, 43] for the quantitative evaluation. Specifically, we use the Fréchet Inception Distance (FID, lower is better) [8], which quantifies the distance between distributions for the real and generated images. FID evaluates the realness of synthesized images. Following [8, 41], we calculate FID for the models trained with limited datasets using 50k generated images and all the real images in the original datasets. We also apply the Inception Score (IS, higher is better) [51]. IS considers the clarity and diversity of generated images.

Implementation details. We choose the state-of-the-art StyleGAN2 [4] as the backbone to verify the effectiveness of APA on limited data. We use the default setups of APA provided in Section 6.2.1 unless specified otherwise. For a fair and controllable comparison, we reimplement all baselines and run the experiments

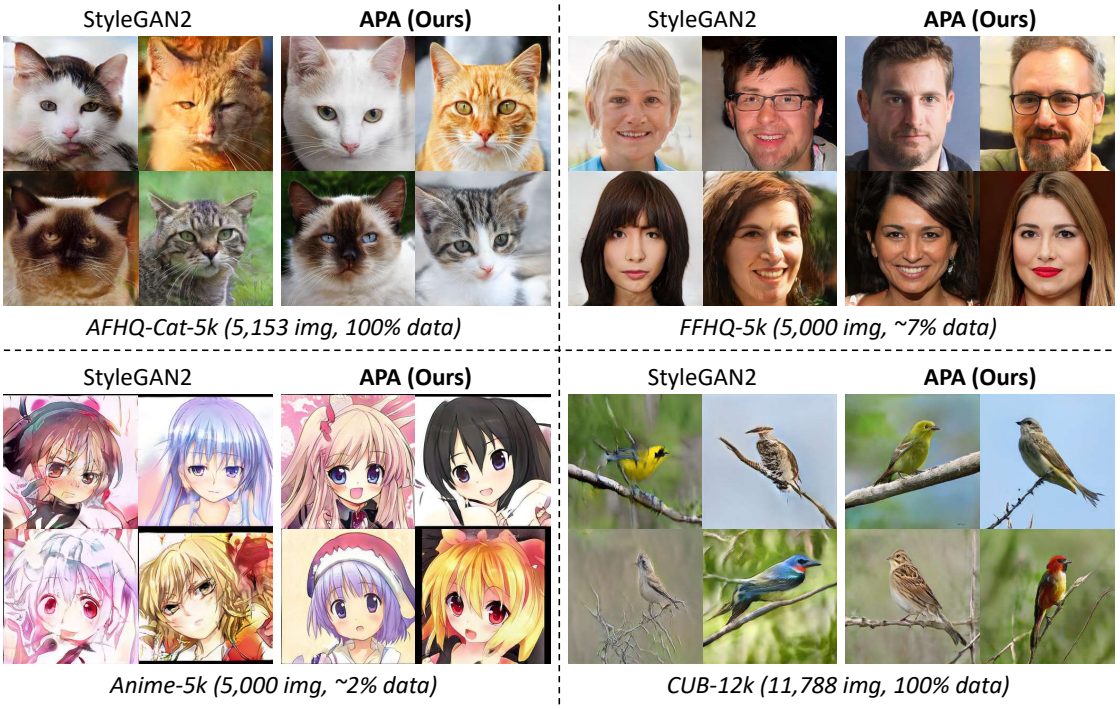

FIGURE 6.4: The proposed APA improves StyleGAN2 [4] synthesized results (256×256, no truncation) on **various datasets** with limited data amounts. We randomly select subsets to confine the size of large datasets (*i.e.*, FFHQ-5k [5] and Anime-5k [7]) and directly use small datasets (*i.e.*, AFHQ-Cat-5k [6] and CUB-12k [9]) whose data amount is already insufficient for StyleGAN2.

TABLE 6.1: The FID (lower is better) and IS (higher is better) scores (256×256) of our method compared to state-of-the-art StyleGAN2 on **various datasets** with limited data amounts.

	AFHQ-Cat-5k		FFHQ-5k		Anime-5k		CUB-12k	
Method	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑
StyleGAN2 [4]	7.737	1.825	37.830	4.018	23.778	2.289	23.437	5.812
APA (Ours)	**4.876**	**2.156**	**13.249**	**4.487**	**13.089**	**2.330**	**12.889**	**5.869**

from scratch using official code. All the models are trained on 8 NVIDIA Tesla V100 GPUs.

6.3.1 The Effectiveness of APA

Effectiveness on various datasets. The comparative results of StyleGAN2 on various datasets with limited data amounts are shown in Figure 6.4. The quality of images synthesized by StyleGAN2 deteriorates under limited data. Ripple artifacts appear on the cat faces and human faces, and the facial features of the anime faces are misplaced. On the bird dataset with heavy background clutter, the generated

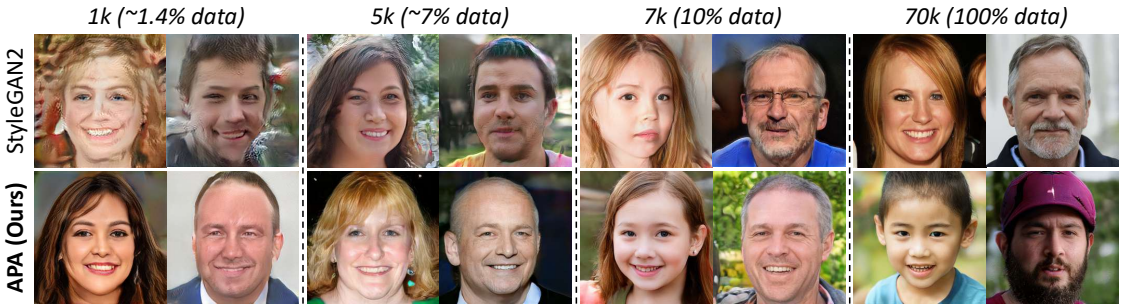

FIGURE 6.5: The effectiveness of APA to improve StyleGAN2 [4] synthesized results (256 × 256, no truncation) on the subsets of FFHQ [5] with **different data amounts**.

TABLE 6.2: The FID (lower is better) and IS (higher is better) scores (256 × 256) of our method on StyleGAN2 trained using the subsets of FFHQ [5] with **different data amounts**.

	1k (∼ 1.4%)		5k (∼ 7%)		7k (10%)		70k (100%)	
Method	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑
StyleGAN2 [4]	86.407	2.806	37.830	4.018	27.738	4.264	3.862	5.243
APA (Ours)	**45.192**	**4.130**	**13.249**	**4.487**	**10.800**	**4.860**	**3.678**	**5.336**

images are completely distorted albeit trained with more data. The proposed APA significantly ameliorates image quality on all these datasets, producing much more photorealistic results. The quantitative evaluation results are reported in Table 6.1. Applying APA contributes to a performance boost of FID and IS in all cases, suggesting that the synthesized images are with higher quality and diversity on different datasets.

Effectiveness given different data amounts. The comparative results on subsets of FFHQ [5], with varying amounts of data, are shown in Figure 6.5. The corresponding quantitative test results are presented in Table 6.2. APA improves the image quality and metric performance in all cases. Notably, the quality of synthesized images by APA on 5k/7k data is visually close to StyleGAN2 results on the full dataset while with an order of magnitude fewer training samples. As for the quantitative results, the IS score of APA on 1k data is better than that of StyleGAN2 on 5k data, and both metrics of APA on 5k data outperform StyleGAN2 results on 7k data. APA can even improve StyleGAN2 performance on the full dataset, further indicating its potential.

Overfitting and convergence analysis. As shown in Figure 6.6, the divergence of StyleGAN2 discriminator predictions can be effectively restricted on FFHQ-7k (10% of full data) by applying the proposed APA. The curves of APA on FFHQ-7k

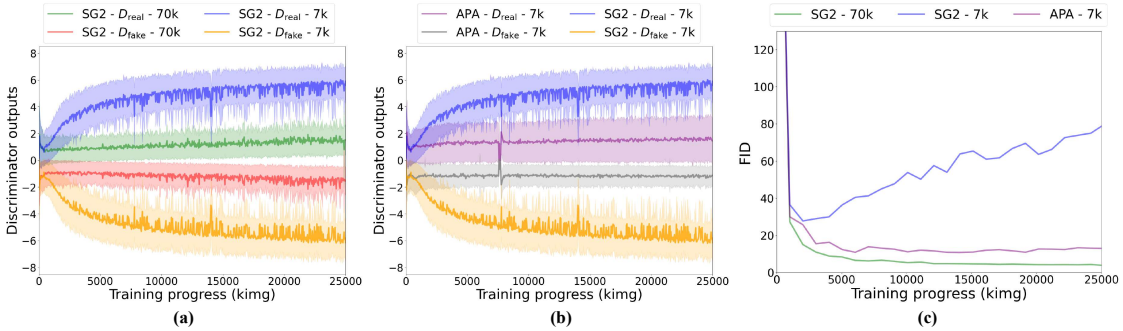

FIGURE 6.6: The **overfitting and convergence status** of APA compared to StyleGAN2 (SG2) on FFHQ [5] (256×256). (a) The discriminator raw output logits of StyleGAN2 on the full (70k) or limited (7k) datasets. (b) The discriminator raw output logits of StyleGAN2 and APA on the limited (7k) dataset. (c) The training convergence shown by FID.

become closer to that of StyleGAN2 on FFHQ-70k, suggesting the effectiveness of APA in curbing the overfitting of the discriminator. Besides, APA improves the training convergence of StyleGAN2 on limited data, shown by the FID curves.

6.3.2 Comparison with Other Solutions for GAN Training with Limited Data

Performance and compatibility. We compare the proposed APA with representative approaches designed for the low-data regime, including ADA [41] and LC-regularization (LC-Reg) [43], which perform standard data augmentations and model regularization, respectively. The results are reported in Table 6.3. As a single method, APA outperforms previous solutions in most cases, effectively improving the StyleGAN2 baseline on both the limited and full datasets. Although ADA [41] achieves slightly better results than our method on FFHQ-5k, it yields a worse FID score with StyleGAN2 on the full dataset. Applying LC-Reg [43] needs careful manual tuning, and its own effect on limited data is not apparent compared to other methods.

It is noteworthy that APA is also complementary to existing methods based on standard data augmentations, *e.g.*, ADA [41]. As can be observed in Table 6.3, APA can further boost the performance of StyleGAN2 with ADA [41] given limited training data, suggesting the compatibility of our approach with standard data augmentations. Combining ADA [41] and APA on FFHQ full data outperforms

TABLE 6.3: The FID (lower is better) and IS (higher is better) scores (256×256) of our method **compared to other state-of-the-art solutions designed for GAN training with limited data** on StyleGAN2. The bold number indicates the best value, and the underline marks the second best.

Method	AFHQ-Cat-5k		FFHQ-5k		FFHQ-70k (full)	
	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑
StyleGAN2 [4]	7.737	1.825	37.830	4.018	<u>3.862</u>	5.243
ADA [41]	6.053	2.119	<u>11.409</u>	<u>4.721</u>	4.018	<u>5.329</u>
LC-Reg [43]	6.699	1.943	35.148	3.926	3.933	5.312
APA (Ours)	<u>4.876</u>	<u>2.156</u>	13.249	4.487	**3.678**	**5.336**
ADA + APA (Ours)	**4.377**	**2.169**	**8.379**	**4.849**	3.811	5.321

TABLE 6.4: The FID (lower is better) and IS (higher is better) scores (256×256) of our method **compared to previous techniques for regularizing GANs** on StyleGAN2 trained with FFHQ-5k [5].

Metric	StyleGAN2 [4]	Instance noise [144]	One-sided LS [51]	APA (Ours)
FID ↓	37.830	40.981	33.978	**13.249**
IS ↑	4.018	4.231	4.029	**4.487**

StyleGAN2 but is slightly inferior to applying APA solely. The degraded performance is mainly affected by ADA [41], which we empirically found might slightly harm the performance when the training data is sufficient. Overall, these methods are not dedicated to improving performance under sufficient data. Nevertheless, as a beneficial side effect, the proposed APA may have this potential.

Training cost. We compare the computational cost of APA against ADA [41], using the same basic official codebase. There is no parameter or memory increment for both methods. As for the time consumption, we test the training cost on 8 NVIDIA Tesla V100 GPUs. On FFHQ-5k (256×256), the average training time of the StyleGAN2 [4] baseline is (4.740±0.100) sec/kimg (*i.e.*, seconds per thousand of images shown to the discriminator). The cost of our method is negligible, slightly increasing this value to (4.789 ± 0.078) sec/kimg. As a reference, the value for ADA [41] is (5.327±0.116) sec/kimg, spending additional time for applying external augmentations.

6.3.3 Comparison with Previous Techniques for Regularizing GANs

As mentioned in Section 2.4, APA is closely related to previous techniques for regularizing GANs. The comparative results on APA and some representative conventional techniques, *i.e.*, instance noise [144] and one-sided label smoothing (LS) [51], are shown in Table 6.4. Applying instance noise [144] may not boost the performance of StyleGAN2 much under limited data. One-sided label smoothing (LS) [51] (with the real label of 0.9) outperforms StyleGAN2 but still has a huge performance gap with our method. This further suggests the effectiveness and usefulness of APA.

6.3.4 Ablation Studies

Ablation studies on variants of APA. We study three key elements of APA, *i.e.*, the overfitting heuristic λ, the deception strength p, and the deception strategy. The version used in our main experiments is denoted as the "main" version, where $\lambda = \lambda_r$, and p is adjusted adaptively. Besides, the "main" version applies the deception strategy that is analogous to one-sided label flipping. As reported in Table 6.5, when using other variants of λ we suggested in Eq. (6.1) (*i.e.*, $\lambda = \lambda_f$ and $\lambda = \lambda_{rf}$), the models achieve comparable performance as the main version. The FID score becomes even better for $\lambda = \lambda_{rf}$, indicating the flexibility of our provided heuristics to extend and modify APA. More interestingly, even a fixed moderate deception probability (*e.g.*, $p = 0.5$) can still work much better than original StyleGAN2 on limited data, albeit slightly inferior to the adaptively adjusted p. This implies the importance of the pseudo augmentation, and the adaptive control scheme can further boost performance without manual tuning. As for the deception strategy, we empirically observe that two-sided label flipping can still outperform StyleGAN2 but is inferior to the main version.

Ablation studies on the threshold t. We further provide the ablation studies on the threshold value t in Table 6.6. The version used in our main experiments is denoted as the "main" version, where $t = 0.6$. It can be seen that the models with different values of t achieve comparable results, outperforming the StyleGAN2 baseline. On FFHQ-5k (256×256), $t = 0.6$ could be a more plausible choice. For a

TABLE 6.5: **Ablation studies on variants of APA** on FFHQ-5k [5] (256×256). We study three key elements of APA, *i.e.*, the overfitting heuristic λ, the deception strength p, and the deception strategy. The "main" denotes the main version used in our previous experiments (*i.e.*, $\lambda = \lambda_r$, p is adaptively adjusted, and the deception strategy is analogous to one-sided label flipping).

Metric	StyleGAN2	main	$\lambda = \lambda_f$	$\lambda = \lambda_{rf}$	$p = 0.5$ (fixed)	two-sided
FID ↓	37.830	<u>13.249</u>	13.470	**12.679**	14.632	15.440
IS ↑	4.018	**4.487**	<u>4.420</u>	4.412	4.403	4.167

TABLE 6.6: **Ablation studies on the threshold t** on FFHQ-5k [5] (256×256). The "main" denotes the main version used in our previous experiments (*i.e.*, $t = 0.6$).

Metric	StyleGAN2	$t = 0.4$	$t = 0.6$ (main)	$t = 0.8$
FID ↓	37.830	13.687	**13.249**	13.984
IS ↑	4.018	4.418	**4.487**	4.395

further explanation, we use $t = 0.6$ as the default value since it works well in most cases. In practice, the value of t could be further adjusted to achieve even better results. Empirically, a smaller t can be chosen when one has fewer data. This means the deception strength p can be adjusted to increase more rapidly since the discriminator is more prone to overfitting when the data amount is fewer.

6.4 Discussion

We have shown the effectiveness of the proposed APA for state-of-the-art GAN training with limited data empirically. With negligible computational cost, APA achieves comparable or even better performance than other types of solutions on various datasets. APA is also complementary to existing methods based on standard data augmentations.

Limitations. Despite promising results, the quality of synthesized images by APA on the datasets with extremely limited data amount (*e.g.*, hundreds of images) can still be improved. Besides, on certain datasets such as FFHQ-5k, applying APA solely may be slightly inferior to approaches based on standard data augmentations. Since we do not apply any external augmentations, these two limitations are both due to the insufficiency of the dataset's intrinsic diversity. These limitations may be approached in the future in two ways: 1) Incorporating better standard

data augmentations to APA. 2) Exploring the issue of data insufficiency from the generator aspect, *e.g.*, using a multi-modal generator [178] to enhance diversity. In addition, we only theoretically verified the convergence and rationality of APA. In future work, the theoretical analysis on the effectiveness of APA could be further explored.

Broader impact. On the one hand, the effectiveness of APA with negligible computational cost will benefit the practical deployment of GANs, especially in the low-data regime. APA may also extend the breadth and potential of solutions to training GANs with limited data and benefit downstream tasks, such as conditional synthesis. On the other hand, APA may also bring potential concerns on its capability to ease the higher-quality fake media synthesis using only limited data, as technology is usually a double-edged sword. However, we believe that these concerns can be resolved by developing better media forensics methods and datasets as countermeasures.

Chapter 7

Conclusion and Future Work

This thesis has demonstrated our ongoing efforts to address the remaining underexplored problems in image and video generation via deep learning. We constructed a large-scale facial video dataset to facilitate our research and secure potential countermeasures to the negative impact of synthesized data by designing a better video manipulation method (Chapter 3). Besides, we proposed a unified and versatile framework for various image-to-image translation tasks with a negligible sacrifice of quality, further extending the practical value of our research (Chapter 4). Apart from these interesting applications, we devised a novel frequency-level loss function that directly optimizes generative models in the frequency domain. The loss is complementary to existing spatial losses of diverse baselines varying in categories, network structures, and tasks, ameliorating generation quality through a more fundamental and theoretical study (Chapter 5). In addition, we introduced a simple yet effective strategy for training GANs with limited data while preserving the quality of synthesis, further easing the practical deployment of powerful GANs with negligible computational cost (Chapter 6). Hopefully, our research has moved this field a small step forward. We discuss other relevant topics and envision our potential future work as follows.

High-fidelity and controllable video synthesis. Despite the remarkable success of image synthesis, high-fidelity and controllable video synthesis remains very challenging. Video generation is much more difficult than image generation as the additional temporal dimension brings in significant variations. A generative model should have the capability to learn the plausible physical motion of objects besides

their appearance. Due to the motion variations, even if a model may synthesize a single frame well, the motion consistency between frames demands additional constraints on the generated results. Besides, our human eyes have evolved to be more sensitive to motion and temporal information, especially when the video resolution is high. In addition, video synthesis typically requires a very expensive computational cost.

Although extensive studies [179–183] have been proposed to address this problem, it remains formidable to generate high-resolution videos with favorable quality and motion coherence in a long-term manner. A representative method, MoCoGAN [181], proposes to decompose motion and content into different parts of random vectors. The quality of its results is not good enough, and the resolution is low (mostly 64×64). MoCoGAN-HD [184] frames the higher-resolution video synthesis problem as discovering a trajectory in the latent space of a pre-trained generator. Such a framework, however, models motion in an autoregressive manner, where the computational cost of training is high. Besides, it is difficult to conduct controllable video synthesis, and the temporal coherence can still be improved.

We are interested in devising a more robust framework for high-fidelity and controllable video synthesis. One of our initial ideas is a progressively growing strategy [49] that grows both the spatial and temporal dimensions. Notably, growing in the temporal dimension may maintain motion consistency more easily and reduce computational cost. Better controllability may be achieved by introducing an intermediate latent space like StyleGAN [5]. On the other hand, the encoder-based GAN inversion method can be a starting point. Given the latent of an initial generated frame as guidance, we can modulate the inversion network temporally or apply a new network to seek the changes of the latent for a video. Intuitively, we can employ a strong image generator [4, 185] as the backbone to ensure single-frame quality. The computational cost can be largely reduced by a non-autoregressive design and improved sparse training [186, 187].

Anime style transfer improvement and extension. Our recent work, DualStyleGAN [188], has demonstrated remarkable success in generating high-quality artistic portrait images by transfer learning with limited data. DualStyleGAN performs exemplar-based high-resolution portrait style transfer by characterizing the content and style of a portrait with an intrinsic style path and a new extrinsic style path, respectively. Despite its success, the quality of results on the Anime styles

can still be improved. The challenge lies in a huge domain gap between the Anime and the real images. For example, the portrait eyes in Anime are usually much larger than the actual human eyes. Such a gap requires a large-degree structure transfer, which may not be tackled well by DualStyleGAN. One potential solution is a better fine-tuning strategy based on adaptive feature blending that applies a learned factor to control the degree of stylization. Such controllability may help enable larger-degree structure transfer to better handle Anime styles.

On another note, performing high-resolution portrait style transfer with fewer training data can be an interesting problem. We sometimes can only collect very limited anime style portraits, which requires the model to be more data-efficient. Involving our introduced technique for GAN training with limited data, APA (Chapter 6), or a multi-modal generator [178] into DualStyleGAN may be solutions. Additional improvement can be achieved by some model regularization strategies, such as [43].

In addition, most research in this area focused primarily on human faces [188–191]. Despite their high research value, generating high-quality anime scenes from complex real-world scene images remains underexplored. The challenges of this task lie in the complexity of the scenes, the unique features of anime style, and the lack of high-quality datasets to bridge the domain gap. Despite promising attempts, previous efforts are still incompetent in achieving satisfactory results with consistent semantic preservation, evident stylization, and fine details. Our initial idea is a semi-supervised image-to-image translation framework to address these challenges. We aim to guide the learning with structure-consistent pseudo paired data, simplifying the pure unsupervised setting. The pseudo data can be derived uniquely from a semantic-constrained StyleGAN2 [4] leveraging rich model priors like CLIP [192]. Besides, we plan to collect a high-resolution anime scene dataset to bridge the domain gap and facilitate future research.

Another interesting extension is 3D-aware high-resolution portrait style transfer. Synthesizing high-quality 3D anime portraits with consistent view information has high practical value. To address this problem, we can start by applying a fine-tuning strategy to advanced 3D-aware image GANs, such as pi-GAN [193], StyleNeRF [194], EG3D [195], StyleSDF [196], GRAM [197]. Since their latent spaces inherit some desirable properties of StyleGAN [5], the fine-tuning scheme used in DualStyleGAN [188] may also suit these backbones.

Image and video generation with diffusion models. Recently, a new powerful generative model, the diffusion model (DM) [47, 48], has attracted more and more attention due to its impressive synthesis capability, especially when involved with multi-modal inputs like text. Diffusion models have been explored in different image and video generation tasks, beating GANs and becoming the new state of the art, which significantly push forward the development of Generative AI. Aside from the extensive exploration of GANs, we are also exploring a representative diffusion model, *i.e.*, Latent Diffusion Model (LDM) [198], on which the prominent Stable Diffusion is based. Previous efforts widely explore text-to-image generation using LDM. We are diving deeply into a new image-to-image translation framework based on LDM to reduce the domain gap between real and simulated data. Our initial idea is a novel loss function at the latent feature level in LDM rather than the common noise level to enable the semantic constraint between the generated and content images. The goal of this loss is to apply different types of constraints for content structure presentation and stylization in LDM, so that the unpaired image-to-image translation can be achieved, extending its practical value. We expect the domain gap can be reduced in terms of visual quality, perceptual metrics, and downstream task performance.

Another important topic is improving image editing with diffusion models. Despite the remarkable performance of diffusion models for image editing, modifying a single object in a complex scene image without affecting others to support localized editing is still challenging. Besides, current editing methods based on diffusion models mainly support the appearance and global semantic editing. In practice, users may not only wish to simply edit the appearance or global semantics but also other various aspects, like object shape, position, and size. More controllable and consistent real image editing is highly desirable. In addition, considering the practical application for general users, reducing the computational cost for additional tuning is beneficial. A potential solution is to harness the internal representations of pretrained text-to-image diffusion models in a more nuanced way. The attention maps inside diffusion models encode structural information about object position and shape [199], and the network feature maps maintain coarse appearance when extracted from appropriate layers. Devising different optimization terms or even manually manipulating these model internals could help more controllable editing without additional tuning.

In the future, we plan to devote more efforts to diffusion models. We wish to delve into multi-modal image and video generation with diffusion models, especially for complex scene images. The interaction between the objects should be considered rather than the object itself. When involving additional multiple modalities, ideally the explicit control of generation and editing can be achieved. Some initial explorations could be conducted to improve ControlNet [200] which incorporates a new single modality. Besides, diffusion models have an obvious disadvantage of slow inference speed due to the multiple sampling steps in the autoregressive design. Devising more efficient diffusion models could be very useful. In addition, a good text prompt is indispensable for diffusion models to generate high-quality results. Involving the powerful ChatGPT into such a text prompt selection task may be interesting. From another aspect, we will not give up GANs. Aligning GAN's performance with diffusion models while maintaining its latent space expressiveness/controllability and fast inference speed is meaningful. Last but not least, we will keep exploring new architectures of diffusion models and even a new family of generative models. We believe that image and video generation via deep learning will be constantly evolving and becoming more promising in the future.

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

Awards

- **Singapore Data Science Consortium (SDSC) Dissertation Research Fellowship**, *Singapore Data Science Consortium, 2021.*
- **Best Paper Award** , *NTU AI Research Student Conference (ARSC), 2022*

Publications

- **Liming Jiang**, Ren Li, Wayne Wu, Chen Qian, Chen Change Loy. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. *In CVPR, 2020.*
- **Liming Jiang**, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, Chen Change Loy. TSIT: A Simple and Versatile Framework for Image-to-Image Translation. *In ECCV, 2020 (Spotlight).*
- **Liming Jiang**, Bo Dai, Wayne Wu, Chen Change Loy. Focal Frequency Loss for Image Reconstruction and Synthesis. *In ICCV, 2021.*
- **Liming Jiang**, Bo Dai, Wayne Wu, Chen Change Loy. Deceive D: Adaptive Pseudo Augmentation for GAN Training with Limited Data. *In NeurIPS, 2021.*

[1] The superscript * indicates joint first authors

- **Liming Jiang**, Zhengkui Guo, Wayne Wu, Zhaoyang Liu, Ziwei Liu, Chen Change Loy, *et al.* . DeeperForensics Challenge 2020 on Real-World Face Forgery Detection: Methods and Results. *arXiv preprint 2102.09471, 2021.*

- **Liming Jiang**, Wayne Wu, Chen Qian, and Chen Change Loy. DeepFakes Detection: The DeeperForensics Dataset and Challenge. *Book chapter. In Handbook of Digital Face Manipulation and Detection - From DeepFakes to Morphing Attacks, Springer, 2022.*

- **Liming Jiang***, Yuxin Jiang*, Shuai Yang, Chen Change Loy. Scenimefy: Learning to Craft Anime Scene via Semi-Supervised Image-to-Image Translation. *Under review, ICCV, 2023.*

- Shuai Yang, **Liming Jiang**, Ziwei Liu, Chen Change Loy. VToonify: Controllable High-Resolution Portrait Video Style Transfer. In ACM Transactions on Graphics (SIGGRAPH Asia), 2022.

- Shuai Yang, **Liming Jiang**, Ziwei Liu, Chen Change Loy. Pastiche Master: Exemplar-Based High-Resolution Portrait Style Transfer. *In CVPR, 2022.*

- Shuai Yang, **Liming Jiang**, Ziwei Liu, Chen Change Loy. Unsupervised Image-to-Image Translation with Generative Prior. *In CVPR, 2022.*

- Shuai Yang, **Liming Jiang**, Ziwei Liu, Chen Change Loy. GP-UNIT: Generative Prior for Versatile Unsupervised Image-to-Image Translation. *In TPAMI, 2023.*

- Jianhui Yu*, Hao Zhu*, **Liming Jiang**, Chen Change Loy, Weidong Cai, Wayne Wu. CelebV-Text: A Large-Scale Facial Text-Video Dataset. *In CVPR, 2023.*

- Yanbo Xu*, Yueqin Yin*, **Liming Jiang**, Qianyi Wu, Chengyao Zheng, Chen Change Loy, Bo Dai, Wayne Wu. TransEditor: Transformer-Based Dual-Space GAN for Highly Controllable Facial Editing. *In CVPR, 2022.*

- Kenny T. R. Voo, **Liming Jiang**, Chen Change Loy. Delving into High-Quality Synthetic Face Occlusion Segmentation Datasets. *In CVPR Workshops / ARSC, 2022.*

- Hao Zhu*, Wayne Wu*, Wentao Zhu, **Liming Jiang**, Siwei Tang, Li Zhang, Ziwei Liu, Chen Change Loy. CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. *In ECCV, 2022.*

Bibliography

[1] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006. xxii, 6, 9, 11, 16, 57, 58, 59, 65, 67

[2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint*, arXiv:1312.6114, 2013. xxii, 1, 3, 6, 11, 16, 26, 27, 39, 46, 57, 58, 59, 65, 68

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. xxii, 1, 3, 6, 7, 12, 14, 15, 16, 18, 39, 58, 59, 66, 69

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. xxiii, xxiv, 1, 4, 6, 7, 9, 12, 16, 17, 57, 60, 66, 72, 77, 78, 79, 80, 85, 86, 87, 89, 94, 95

[5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. xxiii, xxiv, xxvii, 6, 7, 12, 16, 17, 47, 72, 78, 79, 80, 85, 86, 87, 88, 89, 91, 94, 95

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. xxiii, xxiv, 7, 18, 78, 85, 86

[7] Gwern Branwen, Anonymous, and Danbooru Community. Danbooru2019 Portraits: A large-scale anime head illustration dataset. `https://www.gwern.net/Crops#danbooru2019-portraits`. Accessed: 2021-04-10. xxiii, xxiv, 78, 85, 86

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. xxiii, 47, 66, 80, 85

[9] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. 2011. xxiv, 85, 86

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 6, 7, 11, 12, 16, 17, 39, 57, 77, 82, 83, 85

[11] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint*, arXiv:1511.06434, 2015. 1

[12] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint*, arXiv:1411.1784, 2014. 1, 6, 12, 15, 16, 17, 39, 42

[13] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2018. 1, 4, 7, 17, 47, 66, 85

[14] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1, 3, 6, 7, 12, 14, 15, 16, 18, 39, 40, 47

[15] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint*, arXiv:1808.06601, 2018. 1, 30

[16] DeepFakes. `https://github.com/deepfakes/faceswap/`. Accessed: 2019-08-16. 1, 2, 6, 11, 12, 16, 19

[17] DeepFaceLab. `https://github.com/iperov/DeepFaceLab/`. Accessed: 2019-08-20. 6, 11, 16

[18] faceswap-GAN. `https://github.com/shaoanlu/faceswap-GAN/`. Accessed: 2019-08-16. 1, 2, 7, 12, 17, 19

[19] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017. 1, 3, 6, 7, 12, 14, 15, 16, 18, 39, 40

[20] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 1, 6, 7, 8, 12, 14, 15, 16, 18, 39, 41, 47, 48, 49, 50, 52

[21] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. TSIT: A simple and versatile framework for image-to-image translation. In *ECCV*, 2020. 1, 6, 7, 8, 12, 16, 18, 39, 66

[22] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021. 1, 7, 18

[23] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 1, 12, 13, 20, 30, 32, 33, 34, 35, 36

[24] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, 2020. 1, 6, 8, 16, 19

[25] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *NeurIPS*, 2014. 1, 6, 11, 16, 28, 39

[26] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with PixelCNN decoders. In *NeurIPS*, 2016. 1, 6, 11, 39, 57

[27] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. DualGAN: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017. 3, 14, 15, 39, 40

[28] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017. 3, 14, 40

[29] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint*, arXiv:1611.02200, 2016. 3, 14, 40

[30] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017. 3, 14, 40

[31] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017. 3, 14, 40

[32] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017. 3, 8, 14, 15, 39, 40, 41, 48, 51, 71, 72

[33] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *CVPR*, 2018. 6, 7, 12, 15, 16, 18, 44, 46, 47, 48, 51, 71, 72

[34] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 3, 4, 9, 15, 16, 40, 44, 45, 46, 47, 48, 50, 51, 52, 57, 59, 66, 70, 71, 72

[35] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *NeurIPS*, 2019. 6, 7, 8, 12, 14, 15, 16, 18, 39, 40, 41, 46, 47, 48, 50, 51

[36] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *ICLR*, 2020. 3

[37] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *ICML*, 2019. 3, 16, 58

[38] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 16, 66

[39] Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020. 3, 16, 58

[40] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017. 4, 7, 17, 18, 77

[41] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. 4, 9, 18, 78, 79, 81, 85, 88, 89

[42] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *NeurIPS*, 2020. 4, 18, 78

[43] Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. In *CVPR*, 2021. 4, 9, 18, 77, 78, 79, 85, 88, 89, 95

[44] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. *arXiv preprint*, arXiv:2006.02595, 2020. 4, 18, 78

[45] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for gan training. *TIP*, 30: 1882–1897, 2021. 4, 18, 78

[46] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018. 6, 11, 57

[47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 96

[48] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 6, 96

[49] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint*, arXiv:1710.10196, 2017. 6, 7, 12, 16, 17, 66, 94

[50] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 6, 7, 12, 14, 16, 17

[51] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016. 7, 17, 18, 47, 66, 85, 89, 90

[52] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018. 7, 17, 18

[53] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*, 2016. 7, 17

[54] Jianhua Lin. Divergence measures based on the Shannon entropy. *TIT*, 37: 145–151, 1991. 7, 9, 17, 79

[55] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 7, 17

[56] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint*, arXiv:1609.03126, 2016. 7, 17

[57] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50: 157–175, 1900. 7, 17

[58] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 7, 17

[59] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of stylegan imagery. *arXiv preprint*, arXiv:2103.17249, 2021. 7, 17

[60] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *ICCV*, 2021. 7, 9, 12, 18, 57

[61] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 7, 14, 15, 18, 39

[62] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In *ECCV*, 2020. 7, 18

[63] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *arXiv preprint*, arXiv: 2101.05278, 2021. 7, 18

[64] Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. In *NeurIPS*, 2019. 8, 14, 15, 39, 41, 48, 49, 50

[65] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *CVPR*, 2018. 8, 14, 15, 39, 41, 48, 51, 71, 72

[66] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 9, 15, 16, 46, 47, 59, 66, 73

[67] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *CVPR*, 2020. 9, 16, 66, 73

[68] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Deceive D: Adaptive Pseudo Augmentation for GAN training with limited data. In *NeurIPS*, 2021. 9, 77

[69] Ivan Petrov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Jian Jiang, Luis RP, Sheng Zhang, Pingyu Wu, et al. DeepFaceLab: A simple, flexible and extensible face swapping framework. *arXiv preprint*, arXiv:2005.05535, 2020. 11

[70] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, 2018. 12

[71] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *ICCV*, 2019. 12

[72] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. FaceShifter: Towards high fidelity and occlusion aware face swapping. In *CVPR*, 2020. 12

[73] Irene Amerini, Lamberto Ballan, Roberto Caldelli, Alberto Del Bimbo, and Giuseppe Serra. A sift-based forensic method for copy–move attack detection and transformation recovery. *TIFS*, 6:1099–1110, 2011. 12

[74] Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. Detecting image splicing in the wild (web). In *ICMEW*, 2015. 12

[75] Paweł Korus and Jiwu Huang. Multi-scale analysis strategies in prnu-based tampering localization. *TIFS*, 12:809–824, 2016. 12

[76] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP*, 2019. 12, 13, 32

[77] FakeAPP. `https://www.fakeapp.com/`. Accessed: 2019-07-25. 12

[78] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint*, arXiv:1812.08685, 2018. 12, 13, 32

[79] Conrad Sanderson. The vidtimit database. Technical report, IDIAP, 2002. 12

[80] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. *arXiv preprint*, 2019. 12, 13, 20, 32, 33, 34, 36

[81] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016. 12

[82] FaceSwap. `https://github.com/MarekKowalski/FaceSwap/`. Accessed: 2019-08-18. 12

[83] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *arXiv preprint*, arXiv:1904.12356, 2019. 12

[84] Google AI Blog. Contributing data to deepfake detection research. `https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html`. Accessed: 2019-09-25. 12, 13, 20, 23, 32

[85] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint*, arXiv:1910.08854, 2019. 12, 13, 20, 23, 32

[86] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *TVCG*, 20:413–425, 2013. 13, 24

[87] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *WIFS*, 2018. 13

[88] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *IH & MMSEC*, 2016.

[89] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017. 13, 34, 35, 36

[90] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *IH & MMSEC*, 2017.

[91] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *TIFS*, 7:868–882, 2012.

[92] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *WIFS*, 2017. 13

[93] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint*, arXiv:1811.00656, 2018. 13

[94] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *WACVW*, 2019.

[95] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint*, arXiv:1906.06876, 2019.

[96] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *CVPRW*, 2017. 13

[97] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 14, 34, 35, 36, 37

[98] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 34, 35, 36, 37, 43

[99] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–1780, 1997. 34, 35, 36, 37

[100] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 34, 35, 36

[101] 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 *ECCV*, 2016. 14, 34, 35, 36

[102] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In *NeurIPS*, 2018. 14

[103] Xiaoming Yu, Xing Cai, Zhenqiang Ying, Thomas Li, and Ge Li. SingleGAN: Image-to-image translation by a single-generator network using multiple generative adversarial learning. In *ACCV*, 2018. 14

[104] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 15, 28, 39, 40, 45

[105] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint*, arXiv:1508.06576, 2015. 15

[106] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. StyleBank: An explicit representation for neural image style transfer. In *CVPR*, 2017.

[107] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint*, arXiv:1610.07629, 2016. 15, 45

[108] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *ICCV*, 2019. 15

[109] Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, and Jimei Yang. Multimodal style transfer via graph cuts. In *ICCV*, 2019. 15

[110] Shuai Yang, Zhangyang Wang, Zhaowen Wang, Ning Xu, Jiaying Liu, and Zongming Guo. Controllable artistic text style transfer via shape-matching GAN. In *ICCV*, 2019.

[111] Tai-Yin Chiu. Understanding generalized whitening and coloring transform for universal style transfer. In *ICCV*, 2019.

[112] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *ICCV*, 2019.

[113] Ming Lu, Hao Zhao, Anbang Yao, Yurong Chen, Feng Xu, and Li Zhang. A closed-form solution to universal style transfer. In *ICCV*, 2019.

[114] Chunjin Song, Zhijie Wu, Yang Zhou, Minglun Gong, and Hui Huang. ET-Net: Error transition network for arbitrary style transfer. In *NeurIPS*, 2019. 15

[115] Mu Cai, Hong Zhang, Huijuan Huang, Qichuan Geng, and Gao Huang. Frequency domain image translation: More photo-realistic, better identity-preserving. *arXiv preprint*, arXiv:2011.13611, 2020. 16

[116] Rinon Gal, Dana Cohen, Amit Bermano, and Daniel Cohen-Or. SWAGAN: A style-based wavelet-driven generative model. *arXiv preprint*, arXiv:2102.06108, 2021.

[117] Steffen Jung and Margret Keuper. Spectral distribution aware image generation. *arXiv preprint*, arXiv:2012.03110, 2020. 16

[118] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint*, arXiv:1901.06523, 2019. 16, 58

[119] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NeurIPS*, 2008. 16

[120] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 16

[121] Lionel Gueguen, Alex Sergeev, Ben Kadlec, Rosanne Liu, and Jason Yosinski. Faster neural networks straight from JPEG. In *NeurIPS*, 2018. 16

[122] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *CVPR*, 2020.

[123] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *ICCV*, 2019.

[124] Wenlin Chen, James Wilson, Stephen Tyree, Kilian Q Weinberger, and Yixin Chen. Compressing convolutional neural networks in the frequency domain. In *KDD*, 2016.

[125] Seungwook Han, Akash Srivastava, Cole Hurwitz, Prasanna Sattigeri, and David D Cox. not-so-biggan: Generating high-fidelity images on a small compute budget. *arXiv preprint*, arXiv:2009.04433, 2020. 16

[126] A Levinskis. Convolutional neural network feature reduction using wavelet transform. *Elektronika ir Elektrotechnika*, 19:61–64, 2013. 16

[127] Yunhe Wang, Chang Xu, Shan You, Dacheng Tao, and Chao Xu. CNNpack: Packing convolutional neural networks in the frequency domain. In *NeurIPS*, 2016. 16

[128] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020. 17, 57

[129] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *WIFS*, 2019.

[130] Yihao Huang, Felix Juefei-Xu, Qing Guo, Xiaofei Xie, Lei Ma, Weikai Miao, Yang Liu, and Geguang Pu. FakeRetouch: Evading deepfakes detection via the guidance of deliberate noise. *arXiv preprint*, arXiv:2009.09213, 2020. 57

[131] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, 2020. 17

[132] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *ICCVW*, 2019. 17

[133] Yunxuan Wei, Shuhang Gu, Yawei Li, and Longcun Jin. Unsupervised real-world image super resolution via domain-distance aware training. *arXiv preprint*, arXiv:2004.01178, 2020. 17

[134] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *CVPR*, 2020. 17

[135] Nalini M Singh, Juan Eugenio Iglesias, Elfar Adalsteinsson, Adrian V Dalca, and Polina Golland. Joint frequency-and image-space learning for fourier imaging. *arXiv preprint*, arXiv:2007.01441, 2020. 17

[136] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *ECCV*, 2020. 17

[137] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 17, 59, 64

[138] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32:1627–1645, 2009. 17, 59, 64

[139] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 17, 59, 64

[140] Yuanbo Xiangli, Yubin Deng, Bo Dai, Chen Change Loy, and Dahua Lin. Real or not real, that is the question. In *ICLR*, 2020. 17

[141] Ryan Webster, Julien Rabin, Loic Simon, and Frédéric Jurie. Detecting overfitting of deep generative networks via latent recovery. In *CVPR*, 2019. 18

[142] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *ICLR*, 2020. 18

[143] Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for gans. *arXiv preprint*, arXiv:2002.04724, 2020. 18

[144] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *arXiv preprint*, arXiv:1610.04490, 2016. 18, 89, 90

[145] Simon Jenni and Paolo Favaro. On stabilizing generative adversarial training with noise. In *CVPR*, 2019. 18

[146] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017. 18

[147] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *NeurIPS*, 2017. 18

[148] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 18, 46, 47

[149] Soumith Chintala. How to train a GAN? NeurIPS workshop on adversarial training. `https://sites.google.com/site/nips2016adversarial/`, 2016. Accessed: 2021-08-03. 18

[150] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018. 23

[151] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 26

[152] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 30

[153] Kwan-Yee Lin and Guangxiang Wang. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *CVPR*, 2018. 31

[154] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015. 31

[155] Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM TOG*, 37:163, 2018. 32

[156] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 43, 45

[157] Jae Hyun Lim and Jong Chul Ye. Geometric GAN. *arXiv preprint*, arXiv:1705.02894, 2017. 46

[158] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint*, arXiv:1805.08318, 2018. 46

[159] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*, arXiv:1412.6980, 2014. 47

[160] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint*, arXiv:1607.08022, 2016. 47

[161] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, arXiv:1409.1556, 2014. 47, 73

[162] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *arXiv preprint*, arXiv:1805.04687, 2018. 47, 51, 52

[163] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 47, 48, 66

[164] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017. 47, 48, 66

[165] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *CVPR*, 2017. 48, 51, 66

[166] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 48, 66

[167] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. TSIT: A simple and versatile framework for image-to-image translation. *arXiv preprint*, arXiv:2007.12072v1, 2020. 51

[168] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *CVPR*, 2020. 57

[169] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *NeurIPS*, 2020. 57

[170] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1:e3, 2016. 57

[171] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 66

[172] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 66

[173] Radim Šára Radim Tyleček. Spatial pattern templates for recognition of objects with regular structure. In *GCPR*, 2013. 66

[174] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 66

[175] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13: 600–612, 2004. 66

[176] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 66

[177] Cornelius Lanczos. *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Governm. Press Office Los Angeles, CA, 1950. 85

[178] Omry Sendik, Dani Lischinski, and Daniel Cohen-Or. Unsupervised multimodal styled content generation. *arXiv preprint*, arXiv:2001.03640, 2020. 92, 95

[179] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NeurIPS*, 2016. 94

[180] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *ICCV*, 2017.

[181] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *CVPR*, 2018. 94

[182] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint*, arXiv:1907.06571, 2019.

[183] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint*, arXiv:2104.10157, 2021. 94

[184] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *ICLR*, 2021. 94

[185] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 94

[186] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *ICLR*, 2022. 94

[187] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. StyleGAN-V: A continuous video generator with the price, image quality and perks of StyleGAN2. In *CVPR*, 2022. 94

[188] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *CVPR*, 2022. 94, 95

[189] Justin NM Pinkney and Doron Adler. Resolution dependent GAN interpolation for controllable image synthesis between domains. *arXiv preprint*, arXiv:2010.05334, 2020.

[190] Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, and Tat-Jen Cham. AgileGAN: stylizing portraits by inversion-consistent transfer learning. *ACM TOG*, 40:1–13, 2021.

[191] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. VToonify: Controllable high-resolution portrait video style transfer. *ACM TOG*, 41: 1–15, 2022. 95

[192] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 95

[193] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-GAN: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021. 95

[194] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. StyleNeRF: A style-based 3d aware generator for high-resolution image synthesis. In *ICLR*, 2022. 95

[195] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. *arXiv preprint*, arXiv:2112.07945, 2021. 95

[196] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-resolution 3d-consistent image and geometry generation. In *CVPR*, 2022. 95

[197] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. GRAM: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, 2022. 95

[198] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 96

[199] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. arXiv:2208.01626, 2022. 96

[200] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint*, arXiv:2302.05543, 2023. 97