공학박사학위논문

Enhancing Discriminative Capability of GAN Discriminators for Complex Image Generation

복잡한 이미지 생성을 위한 GAN 판별기 강화 방법

2023 년 2 월

서울대학교 대학원

컴퓨터 공학부

이 한 빛

Enhancing Discriminative Capability of GAN Discriminators for Complex Image Generation

복잡한 이미지 생성을 위한 GAN 판별기 강화 방법

지도교수 이 상 구

이 논문을 공학박사학위논문으로 제출함

2022 년 11 월

서울대학교 대학원

컴퓨터 공학부

이 한 빛

이한빛의 공학박사 학위논문을 인준함

2022 년 12 월

위 원 장	김건희	(인)
부위원장	이상구	(인)
위 원	황승원	(인)
위 원	심준호	(인)
위 원	박재휘	(인)

Abstract

Generative Adversarial Network (GAN) is one of the most successful generative models in recent years. GAN involves adversarial training between two networks, a generator and a discriminator, which provides a novel and powerful way of modeling high-dimensional data distribution such as images. At the core of this mechanism, the discriminative capability of the discriminator plays a significant role, because the generator can improve itself only to the extent that the discriminator can distinguish between real and fake samples. In this dissertation, we sought for the improvement of GAN models by proposing three methods to enhance the discriminative capability of the discriminator.

To improve conditional image generation for complex multi-label conditions, we propose an attention-based conditional discriminator that allows the discriminator to focus on local regions that are relevant to given labels. In addition, we propose a product-of-Gaussian based latent sampling method to better encode the multi-label condition. Both proposed architectures for discriminator and generator improve the controllability of the image generation process.

We then study discriminator enhancement for more complex data distributions, such as scene images with multiple objects. Due to the high structural complexity of scene images, the discriminator is under heavy burden to distinguish complex structural differences between real and fake scene images. To aid the discriminator, we design a multi-scale contrastive learning task to enhance local representations of the discriminator. The proposed auxiliary task allows us to learn a powerful discriminator that can better incentivize the generator to improve the synthesis quality of scene images.

Finally, we explore a way to utilize pretrained scene understanding models for the discrimination process. Since the pretrained models contain rich knowledge on complex structures of scene images, we propose to use their pretrained representations to relieve the burden of the discriminator. To take full advantage of both common and per-task knowledge available in different pretrained models, we propose to ensemble their features to form a set of unified multi-scale features.

With extensive evaluation and analysis on challenging image domains, we show that the proposed methods achieve meaningful improvement on modeling complex image distributions. We believe these achievements would help increase the utility of GAN models, and facilitate their downstream applications as well.

Keywords: Generative Adversarial Networks, Deep Generative Models, Image Generation, Conditional Image Generation, Discriminator Enhancement, Scene Generation, Self-Supervised Learning, Transfer Learning

Student Number: 2013-20865

Contents

Abstract **1**

1 Introduction **9**

1.1 Deep Generative Models 9

1.2 Generative Adversarial Networks 11

1.3 Scope of Dissertation 12

1.4 Contributions 13

2 Preliminaries and Related Work **15**

2.1 Generative Adversarial Networks 15

2.2 Architectural Improvement 16

2.3 Objective Functions and Regularization 19

2.4 Auxiliary Task 20

2.5 Transfer Learning for GAN 21

2.6 Evaluation of Generative models 22

3 Attention-based Discriminator for Multi-label to Image Generation **24**

3.1 Motivation 24

3.2 Related Work 26
3.3 Method 28
3.3.1 Multi-label Attention for Discriminator 28
3.3.2 Product-of-Gaussian Condition Prior for Generator 30
3.3.3 Visual-Semantic Embedding 31
3.4 Experiment 33
3.4.1 Quantitative Result 35
3.4.2 Qualitative Result 40
3.5 Chapter Summary 44

4 Multi-scale Contrastive Learning for Complex Scene Generation 45
4.1 Motivation 45
4.2 Related Work 48
4.3 Method 50
4.3.1 Multi-scale Discriminator with Multi-level Branches . . . 51
4.3.2 Multi-scale Contrastive Learning for GAN 53
4.3.3 Full Objective 55
4.3.4 Implementation and Training 56
4.4 Experiment 56
4.4.1 Comparison to State-of-the-Art 60
4.4.2 Ablation Study 63
4.4.3 Analysis on Training Dynamics 68
4.5 Chapter Summary 68

5 Leveraging Pretrained Vision Models for Complex Scene Generation 70
5.1 Motivation 70

5.2 Method 72
5.2.1 Leveraging Pretrained Vision Models 72
5.2.2 Feature Ensemble across Scales and across Models 73
5.3 Experiment 74
5.3.1 Comparison Result 78
5.3.2 Ablation Study 80
5.4 Chapter Summary 82

6 Conclusion & Future Work **84**

7 Appendix **88**
7.1 Detailed Network Architecture 88
7.1.1 Network Architecture of ADGAN 88
7.1.2 Network Architecture of MsConD 88
7.2 Additional Samples 89
7.2.1 Comparison of additional samples of MsConD 89

초록 **113**

List of Figures

2.1 Related research map on Generative Adversarial Networks 17

3.1 Illustration of the ADGAN architecture 25
3.2 Illustration of Multi-label attention-based conditional discriminator of ADGAN 29
3.3 Illustration of Product-of-Gaussian based conditional prior sampling of ADGAN 30
3.4 Number of training images for each attribute in Polyvore dataset 33
3.5 Comparison of correctness per attribute 37
3.6 Comparison of correctness per attribute group 38
3.7 Ablation result of correctness per attribute 39
3.8 Ablation result of correctness per attribute group 40
3.9 Qualitative comparison to the baseline 41
3.10 Images generated by ADGAN 42
3.11 Images generated by ADGAN with a set of attributes 43

4.1 Illustration of proposed MsConD 47
4.2 Illustration of proposed discriminator architecture 51

4.3 Illustration of spatially consistent pixel-level contrastive learning in MsConD 53
4.4 Comparison of generated samples 61
4.5 Samples Generated with MsConD 62
4.6 Quantitative Ablation Result 64
4.7 Quantitative ablation result for each object category 65
4.8 Qualitative ablation result on Livingroom dataset 66
4.9 Comparison of training progress on Cityscapes 67

5.1 Overview of the proposed method 72
5.2 Illustration of proposed feature ensemble method 73
5.3 Comparison of generated samples. Red rectangles show imperfect object structures. Blue rectangles show repetitive stains making messy layouts. 79

7.1 Uncurated Samples for Cityscapes 92
7.2 Uncurated Samples for Livingroom 93
7.3 Uncurated Samples for Kitchen 94

List of Tables

3.1 Number of images with different number of associated attributes in Polyvore dataset 32
3.2 Quantitative comparison result 36

4.1 Comparison result on Scene-level generation metrics 58
4.2 Object-level metrics for each object category 59

5.1 Comparison result on scene-level metrics 76
5.2 Object-level metrics for each object category 77
5.3 Ablation result using different pretrained networks 81
5.4 Effectiveness of feature ensemble 82

7.1 Discriminator Architecture of ADGAN 90
7.2 Generator Architecture of ADGAN 90
7.3 Discriminator Architecture of MsConD 91

Chapter 1

Introduction

1.1 Deep Generative Models

Over the past decade, AI systems have achieved remarkable success in a variety of tasks including computer vision, recommendation, and language processing (Zhang et al., 2022). At the heart of this success, there exist a vast amount of data on which AI systems can be trained. The world around us is represented and perceived by various forms of data, such as images, natural language, and speech. One of the important methodologies to fundamentally understand and utilize these data is generative model.

Generative models aim to learn and model how the data itself can be generated. Once the model is successfully trained, we can estimate the likelihood of given data samples or generate new realistic samples. A variety of useful applications are available via generative models since we can sample new data points from the model. For example, users can generate or edit images or texts toward their intention (Isola et al., 2017; Zhu et al., 2017a; Zhang et al., 2017;

He and Deng, 2017; Pang et al., 2021). Generated data can be further utilized to augment the training dataset to improve the prediction models (Antoniou et al., 2017; Shin et al., 2018; Yoo et al., 2019; Zhang et al., 2021). Generative models are also useful for explaining and analyzing other AI models by visualizing the decision boundaries in an intuitive way (Verma et al., 2020; Sauer and Geiger, 2021; Lang et al., 2021).

Despite such powerful features and possibilities, generative models are known to be much more challenging to train compared to analogous discriminative models. Unlike discriminative models, generative models are required to generate all the values that make up a data sample which often lies in high dimensional data space such as images of high resolution. Therefore, estimating the density function of high dimensional data often becomes an intractable task requiring heavy computational resources.

However, in recent years, the development of deep generative models has significantly alleviated the difficulty and has shown impressive results for complex and high-dimensional data distributions. From Deep Belief Networks (Hinton et al., 2006) to Deep Boltzmann Machines (Salakhutdinov and Larochelle, 2010), to Variational Auto-encoders (Kingma and Welling, 2014) and Generative Adversarial Networks (Goodfellow et al., 2014), through the advances on network architectures and density estimation methods, deep generative models have greatly improved the training efficiency by leveraging neural networks as an efficient density estimator. Among the different types of deep generative models, Generative Adversarial Network (GAN) is one of the most successful models within a decade since it can produce much sharper and discrete outputs compared to the results of other models. In this dissertation, we focus on Generative Adversarial Networks and explore ways to improve it.

1.2 Generative Adversarial Networks

GAN consists of two neural networks: a generator and a discriminator. The generator is a network that transforms random noise vectors into synthetic data samples and the discriminator is a classifier that classifies input samples into real or generated ones. The core mechanism of GAN is an adversarial training scheme where two networks are trained to satisfy conflicting objective functions. Concretely, the discriminator is trained to correctly discriminate real data from generated data, while the generator is trained so that the generated data is determined by the discriminator to be real one, that is, to deceive the discriminator.

While GAN models have received much attention for its superior capability, they also have non-trivial limitations and challenges. Since GAN was first proposed, its unstable training has been pointed out as the biggest limitation (Roth et al., 2017; Wu et al., 2020). Training process easily diverges if hyper-parameters are not cautiously tuned and the results easily collapse to few samples, i.e., mode collapse. Another drawback of GAN is its narrow coverage over the data distribution. While recent GAN models work well with relatively simple data distributions such as human faces (Karras et al., 2019, 2020b), they easily fail to achieve the same level of realism for more complex image domains such as scenes with various objects classes (Gadde et al., 2021). Therefore, even state-of-the-art models result in mottled layouts and discontinued semantic structures. In this dissertation, we focus on enhancing the discriminator to overcome these limitations.

In GAN training, a generator and a discriminator rely on each other's performance to develop themselves competitively. The discriminator evolves more precisely as the generated samples become more realistic, and the generator

relies on the discriminator's discriminative ability to improve the fidelity of its output samples. At the heart of this mechanism, the discriminative ability of the discriminator plays a significant role. Since the generator is able to only improve itself to the extent that the discriminator distinguishes between real and fake samples, training a powerful and robust discriminator largely determines the overall generation performance. Therefore, we explore the ways to strengthen the discriminator learning in three different directions: (1) architectural improvement, (2) auxiliary task for discriminator, and (3) use of pretrained vision models.

1.3 Scope of Dissertation

Architectural improvement For most neural models, designing effective building blocks of neural networks is a key element to improve the training efficiency and performance. Likewise, the synthesis quality of GAN is also highly dependent on how we design the network architecture for both the generator and discriminator. In Chapter 3, we describe Attention-based Discriminator for Conditional GAN (Lee and Lee, 2019), which leverages attention mechanism to strengthen the discriminator and improves conditional generation performance for complex multi-label condition. In addition, we propose a prior latent conditioning method using the product-of-Gaussian to encode a set of labels and generate diverse and accurate samples.

Auxiliary task for Discriminator Several recent studies have shown that GAN training can be improved by assigning auxiliary tasks to the discriminator in addition to the original binary classification task (Zhang et al., 2020; Jeong and Shin, 2021; Yang et al., 2021). In Chapter 4, based on these findings, we propose MsConD (Lee et al., 2023) which leverages multi-scale contrastive learn-

ing as an auxiliary task for the purpose of improving complex scene generation. MsConD uses a multi-scale discriminator that performs patch-level real/fake classification to improve fidelity of local semantic structures in the scene image. To enhance the multi-scale discriminator, we propose to assign contrastive learning tasks for local representations in multiple scales. By jointly optimizing both original classification task and contrastive learning task, the discriminator can enhance its discriminative ability to better incentivize generator to improve its generation performance.

Leveraging Pretrained Vision models It has now become a ubiquitous process to boost the performance of downstream tasks by transferring general-purpose representations learned on large-scale datasets. Recently, several studies have shown that pretrained representations can benefit discrimination process in GAN to improve the generation quality as well as the convergence speed (Sauer et al., 2021; Kumari et al., 2022). Chapter 5 describes FEGAN, which leverages pretrained scene understanding models to improve complex scene generation. FEGAN employs powerful pretrained models trained for various scene understanding tasks such as object detection, semantic segmentation, and depth estimation, and ensembles their multi-scale features to aid the discriminator. Proposed ensemble scheme fully leverages knowledge learned by different scene understanding models to help GAN better synthesize complex scenes.

1.4 Contributions

Contributions of this dissertation can be summarized as three-fold:

- We propose an advanced architecture for generative adversarial networks in context of conditional image synthesis by utilizing attention mecha-

nism for conditional discriminator and product-of-Gaussian based condition aggregation for conditional generator. Our method provides improved controllability on image synthesis for complex multi-attribute conditions.

- We propose an auxiliary task which is designed to improve the discriminative capability of discriminator on complex scene images containing multiple objects. By leveraging self-supervised learning scheme, our method can improve synthesis quality of generator without any other prior information on complex scenes, such as object-level or pixel-level labels.

- We also explore the utility of pretrained vision models trained on various scene understanding tasks for scene image generation. We propose a feature-level ensemble method which allows the discriminator to effectively utilize pretrained representations from multiple models, thereby further improve the synthesis quality of complex scene images.

Chapter 2

Preliminaries and Related Work

In this chapter, we introduce Generative Adversarial Networks (GANs) with formal notations that will be used in later chapters. Then we review the related studies that contributed to GAN models, especially to enhance the discriminative ability of the discriminator. Related works are classified into four major categories and reviewed in detail. Figure 2.1 shows the overview of related work on GAN research.

2.1 Generative Adversarial Networks

A standard GAN involves a minimax optimization between two networks, a generator G and a discriminator D as follows:

$$\min_G \max_D \mathcal{L}_{adv}(G, D) = \mathbb{E}_{x \sim p_{data}} \left[\log D(x)\right] + \mathbb{E}_{z \sim p_z} \left[\log(1 - D(G(z)))\right], \tag{2.1}$$

where p_{data} is an empirical data distribution where a real data sample x is sampled and p_z is a known prior distribution where a noise z is sampled. G generates a fake image $G(z)$ from a sampled noise z and D computes the prediction score for both the real image x and the generated image $G(z)$. The adversarial training encourages D to correctly distinguish real images from generated images while G to synthesize realistic-looking images so that the generated images can be distinguished as real ones by D.

While the original GAN formulation only considers unconditional learning of data distributions, follow-studies (Mirza and Osindero, 2014a; Odena et al., 2017; Miyato and Koyama, 2018) have developed its conditional variants those enable conditional data generation. Given p_{data} as an empirical data distribution of labeled data, i.e., (x, y), the conditional GAN can be formulated as follows:

$$\min_G \max_D \mathcal{L}_{adv}(G, D) = \\ \mathbb{E}_{(x,y)\sim p_{data}} \left[\log D(x|y)\right] + \mathbb{E}_{z\sim p_z} \left[\log(1 - D(G(z|y)))\right], \tag{2.2}$$

where y is a condition corresponds to a real sample x. As can be seen, the conditional models take the condition y as well as x to train G and D.

2.2 Architectural Improvement

The GAN models, like all other deep neural network-based models, are largely affected by the network architecture in training efficiency and generation performance. The seminar paper (Goodfellow et al., 2014) uses simple multi-layer perceptrons for both generator and discriminator, therefore the model was validated only on simple images such as MNIST (Deng, 2012). DCGAN (Radford et al., 2015) has first explored the use of convolutional networks for GAN

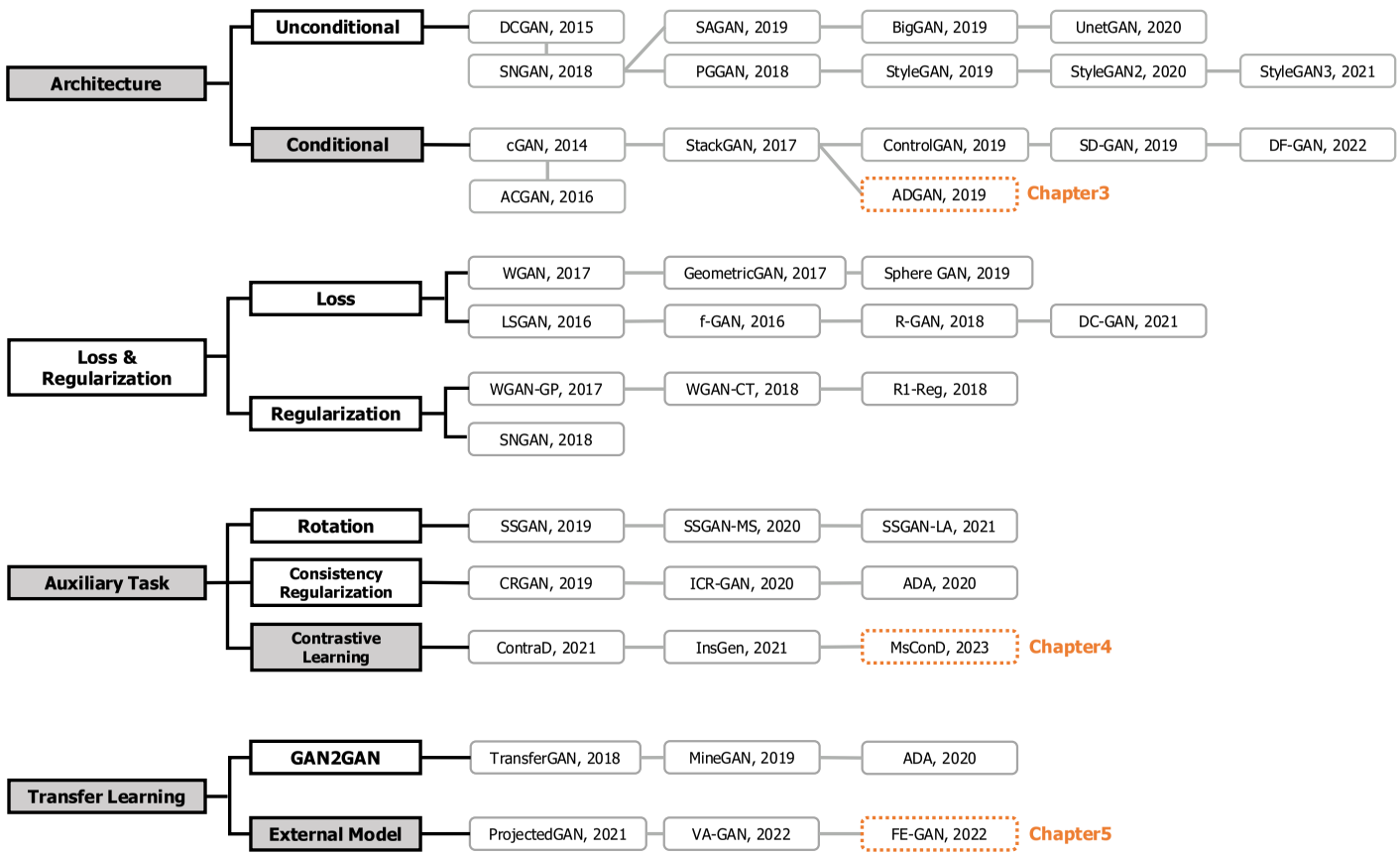


Figure 2.1 Related research map on Generative Adversarial Networks

framework and succeeded in generating more complex images such as bedroom images. The network architectures are further improved by employing residual networks (He et al., 2016b; Miyato et al., 2018) and self-attention layers (Zhang et al., 2019; Brock et al., 2019). While most GAN models utilize symmetric structures for generator and discriminator, a notable exception is recently proposed UnetGAN (Schonfeld et al., 2020) which employs Unet (Ronneberger et al., 2015) architecture only for the discriminator to improve discriminative ability and has shown improved synthesis result.

While major architectural improvement has been achieved and validated on unconditional GAN models, there also have been significant advances on conditional GAN models. Conditional GANs aim to generate images for given various types of condition such as class label, text, or other images. For class label condition, cGAN (Mirza and Osindero, 2014a) is the first model that successfully extend GAN to its conditional variant. Conditional mechanism is further improved by introducing auxiliary classifier (Odena et al., 2017) and projection based discriminator (Miyato and Koyama, 2018). For text condition, StackGAN (Zhang et al., 2017) is the first text-to-image GAN model which utilizes a text encoder to map a condition text into corresponding sentence embedding. After StackGAN, there have been various attempts to improve the controllability using word-level spatial attention (Li et al., 2019), siamese discriminator (Yin et al., 2019), and matching-aware gradient penalty (Tao et al., 2022). Image-to-image translation is the generation task which utilizes images as the condition. Pix2Pix (Isola et al., 2017) and CycleGAN (Zhu et al., 2017a) are the representative models for the image-to-image translation task. Text-to-image and image-to-image synthesis tasks have been extensively studied since these seminar works until recently (He and Deng, 2017; Pang et al., 2021).

In Chapter 3, we propose ADGAN which is one of the first approaches that

introduced the attention mechanism to GAN models like SAGAN (Zhang et al., 2019). However, different from SAGAN, ADGAN proposes a attention-based discriminator for conditional tasks rather than using self-attention blocks.

2.3 Objective Functions and Regularization

One of the major drawbacks of GAN is its unstable training (Mescheder et al., 2017, 2018a). To stabilize the training of GANs, alternative objective functions as well as regularization techniques have been extensively studied. The objective function of original GAN is formulated as the binary classification task between real and fake samples, which is equivalent to minimize the Jensen-Shannon Divergence between the real and fake data distributions. Wasserstein GAN (Arjovsky et al., 2017) and its improved version (Gulrajani et al., 2017) propose to minimize the Earth-Mover distance to further stabilize the learning process. Hinge loss (Lim and Ye, 2017b) is another widely adopted loss function for GAN, which are widely adopted in various conditional generation tasks. More recently, dual-contrastive GAN (Yu et al., 2021) has shown contrastive loss can be utilized to improve the discriminative ability between real and fake images.

Recent studies show that powerful regularization techniques are more important than the alternative loss functions. One of the representative approaches is to penalize the norm of the gradient for the discriminator input during the training process (Gulrajani et al., 2017; Mescheder et al., 2018b). The gradient penalty (Arjovsky et al., 2017; Gulrajani et al., 2017) imposes a Lipschitz constraint on the discriminator and helps convergence. Another notable approach is spectral normalization which was introduced in SNGAN (Miyato et al., 2018). It controls the Lipschitz constant of the discriminator by controlling the spectral norm of the layer weights. Unlike gradient penalty, spectral normalization

does not require to additionally calculate the regularization term, therefore it enables much efficient training. While a variety of regularization approaches (Webster et al., 2019; Yang et al., 2019; Tseng et al., 2021; Ni et al., 2022) have been proposed till recently, gradient penalty and spectral normalization have been utilized as the most popular and effective methods.

2.4 Auxiliary Task

A group of works (Chen et al., 2019; Tran et al., 2019; Hou et al., 2021) have shown that the rotation prediction task prevents catastrophic forgetting in GAN and leads to better results. Consistency regularization (Zhang et al., 2020; Zhao et al., 2021) stabilizes GAN training by imposing consistency of discriminator output between a clean image and its augmented version. More recently, several studies have explored the use of the instance discrimination task (Wu et al., 2018; He et al., 2020; Chen et al., 2020) as an auxiliary task to further enhance the discriminator (Zhao et al., 2020b; Jeong and Shin, 2021; Yang et al., 2021). The self-supervised pretext tasks generally involve various image transformation functions to acquire different views of an image. In GAN training, differentiable image transformations (Karras et al., 2020a; Zhao et al., 2020a) applied on both real and fake images have shown to stabilize the training in limited data regimes and improve the data efficiency.

While the main optimization task of GAN is the binary classification task for real and generated samples, recent studies have shown that a various self-supervised learning tasks can aid the real/fake classification task and help to train a more robust discriminator. A group of works (Chen et al., 2019; Tran et al., 2019; Hou et al., 2021) have shown that the rotation prediction task prevents catastrophic forgetting in GAN and leads to better results. Consistency

regularization (Zhang et al., 2020; Zhao et al., 2021) stabilizes GAN training by imposing consistency of discriminator output between a clean image and its augmented version. More recently, several studies have explored the use of the instance discrimination task (Wu et al., 2018; He et al., 2020; Chen et al., 2020) as an auxiliary task to further enhance the discriminator (Zhao et al., 2020b; Jeong and Shin, 2021; Yang et al., 2021). The self-supervised pretext tasks generally involve various image transformation functions to acquire different views of an image. In GAN training, differentiable image transformations (Karras et al., 2020a; Zhao et al., 2020a) applied on both real and fake images have shown to stabilize the training in limited data regimes and improve the data efficiency.

Previous studies mainly focus on self-supervised tasks defined on global representations. This is effective for image domains consisting of a single object class, but the improvement would be limited for more complex scene images. In Chapter 4, we propose a self-supervised task designed to enhance local representations for multiple scales and assign it to the discriminator to increase global-to-local fidelity of generated scene images.

2.5 Transfer Learning for GAN

Most GAN models are trained from scratch with randomly initialized parameters, thus it takes a long training time to converge. In addition, if the amount of training data is relatively small, training may be easily unstable and the synthesis quality significantly decreases. To alleviate these difficulties, transfer learning approaches designed for GAN have been proposed. Most of these approaches aim to transfer and fine-tune the generator parameters from data-sufficient to data-scarce domains.

More recently, a line of works have shown that the discriminator can also benefit from pretrained networks yielding better synthesis quality and training efficiency (Sauer et al., 2021; Kumari et al., 2022). ProjectedGAN (Sauer et al., 2021) makes use of EfficientNet (Tan and Le, 2019) trained on ImageNet classification task to extract image representations. Light-weight discriminators are trained to distinguish between real/fake images based on the extracted representations. Their scheme greatly improves the convergence speed as well as synthesis quality. VAGAN (Kumari et al., 2022) also utilizes fixed pretrained vision models to extract discriminative features and shows that their approach can improve generation performance when training data is relatively scarce.

In Chapter 5, we extend these findings to explore how to use various scene understanding models to improve the generation performance of complex scene images. Unlike previous works, we propose an feature ensemble method to fully take advantage of multi-scale representations extracted from multiple pretrained models.

2.6 Evaluation of Generative models

Evaluating the quality of trained generative models is crucial but non-trivial task. While a variety of works rely on human evaluation to comparatively assess their models, human evaluation requires costly labor. Moreover, visual inspection and assessment can easily be subjective when performed with a small group of evaluators. Therefore, researchers have sought for reliable automatic quantitative metrics. The one of the earliest attempts is Inception score (Salimans et al., 2016). Inception score assess the quality of generated images with Inception network (Szegedy et al., 2016) trained on ImageNet (Deng et al., 2009). The score is calculated by multiplying sharpness and diversity score measured with

model predictions of generated images. While inception score serves as a reasonable metric, its quality is largely limited when the generated images do not belong to ImageNet classes. Frechet inception score (Heusel et al., 2017a) and Kernel inception score (Bińkowski et al., 2018) aim to provide better metrics by utilizing not only the generated images but also the real images. They measure the discrepancy between real and generated distributions with Frechet distance and kernel Maximum Mean Discrepancy, respectively. They systematically show that the proposed metrics are precisely aligned with perceptual quality assessed by human evaluators. Precision and Recall (Kynkäänniemi et al., 2019) are another popularly used metrics. Precision measures the ratio of generated samples that fall into real sample distribution and Recall measures the ratio of real samples that fall into generated sample distribution. They use k nearest neighbors in the counterpart sample set to identify pseudo-identical samples to calculate the metrics.

Chapter 3

Attention-based Discriminator for Multi-label to Image Generation

3.1 Motivation

Conditional image generation aims to synthesize novel realistic images those reflect user-specified conditions. These conditions often include sets of multiple attributes that match the complex user intention. For example, imagine a fashion designer who are designing new clothes. He or she might want new reference images that follow the target design concept such as "a white flare dress with floral pattern".

We formulate this task as a multi-label to image synthesis. A concept or condition the user has in mind can be represented as a set of attribute labels. For above example, the input condition can be described as a attribute set $\{white, flare, dress, floral\}$. Therefore, our aim is to generate synthetic images those meet a set of given multiple attributes.

Recently as Generative Adversarial Networks (Goodfellow et al., 2014) have

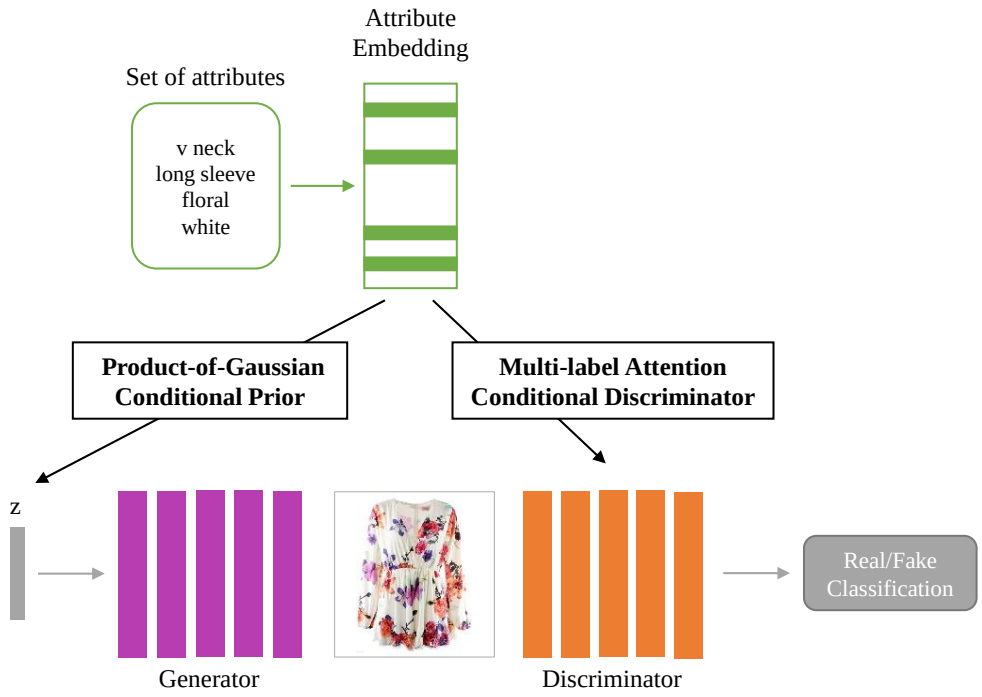


Figure 3.1 Illustration of the ADGAN architecture

shown impressive results in synthesizing realistic images, a bunch of models based on conditional GAN (Gauthier, 2014; Mirza and Osindero, 2014b) have been proposed. Nevertheless, most approaches deal with small number of attributes (van den Oord et al., 2016; Yan et al., 2016) or a single class label (Odena et al., 2016). Therefore, the existing methods struggle for more complex conditional generation task involving large number of attribute labels such as fashion images. In addition, the existing methods have been only studied on image domains where the attributes are densely annotated with extensive human labor, such as CelebA dataset (Liu et al., 2015). But for many real world image domains, the attributes are rather sparse and fuzzy.

To build a image generation model for more complex conditions, we imple-

ment three modules that can be integrated to conditional GAN model. First, we construct a visual-semantic embedding model to obtain meaningful attribute embeddings by relating the visual feature from image and associated label sets, so the attribute embeddings contain rich semantic representations. We then design a novel conditional modules for both generator and discriminator to improve the conditional GAN model. Two modules are *multi-label attention-based discriminator* and *product-of-Gaussian generator*. Since each attribute is associated to different spatial regions in an image, we propose to make discriminator to pay attention to relevant attributes when discriminating an input image. In addition, we design a new conditional prior based on product-of-Gaussian for the generator. The intuition under this module is that we can encode the combination of attributes in a principle and efficient way by sampling condition vector from product of multiple Gaussian distributions derived from each attribute in a given set.

We validate our method on a real world fashion dataset. For quantitative evaluation, we suggest to use a pre-trained attribute classifier to evaluate the generated images with correctness and coverage metrics. The result shows that our model significantly outperforms the baseline model showing better controllability over generated images. We also present generated samples of our model and compare with the samples of baseline model to show efficacy of the proposed method.

3.2 Related Work

Conditional Generative Adversarial Networks. There has been efforts to control the image generation with certain condition. As the research on GAN has progressed, the methods about adding extra information to control genera-

tion have been studied. cGAN (Mirza and Osindero, 2014b) is implemented by supplying both generator and discriminator with class labels to learn conditional distribution. Odena et al. (2016) suggests to use discriminator as an auxiliary classifier to output the predicted class labels. van den Oord et al. (2016) performs attribute-controlled image generation using PixelCNN. These approaches use condition of a class label or small number of attributes. There are a line of research studying image synthesis from unstructured text. Reed et al. (2016) uses conditional PixelCNN to generate images with text descriptions. Zhang et al. (2017) propose a two-stage training strategy that can generates images with higher resolution.

Attention-based architectures. The attention mechanism has become an essential part of models for machine translation (Bahdanau et al., 2014), image captioning (Xu et al., 2015), and visual question answering (Yang et al., 2016). However attention mechanism has not much been explored in the context of GAN. Xu et al. (2017) propose to use attention for multi-stage generator for text-to-image synthesis. Zhang et al. (2019) exploit self-attention mechanism to unsupervised GAN to improve the image generation. To our best knowledge, the proposed ADGAN apply attention mechanism to discriminator in the context of conditional image generation.

Generative models for fashion domain. As there has been remarkable progress in deep generative model including GAN, research on generating images for fashion application has begun to start. Zhu et al. (2017b) build a GAN based model that generates images for imagination of wearing image. Jiang and Fu (2017) propose a style generator which generates a clothing image with given input patterns on it. Sbai et al. (2018) propose a creative loss function to

generate novel clothing images.

3.3 Method

Our goal is to train a generator G that can generate a realistic image $G(z, y)$ consistent with a given set of attributes $y = \{y_1, y_2, ..., y_n\}$. The attributes are represented as fixed-dimensional vectors where the vectors are pre-trained visual-semantic embeddings described in Section 3.3.3 . As shown in Figure 3.1, the proposed Attention-based Discriminator Generative Adversarial Network (ADGAN) has a novel discriminator and a generator structure in context of multi-attribute to image synthesis. Attention-based discriminator is optimized to discriminate real from synthesized images with locally attended features of the given attribute set. For the generator, ADGAN encodes a set of attribute vectors to a single condition vector sampled from Gaussian which is the product of multiple Gaussian distributions derived from each attribute vector.

3.3.1 Multi-label Attention for Discriminator

Figure 3.2 shows discriminator design with multi-label attention. The proposed ADGAN integrate attention mechanism for discriminator D to compute probability $D(x, y)$ from an input image x and the corresponding set of attributes $y = \{y_1, y_2, ..., y_n\}$. The discriminator is a convolutional neural network composed of several downsampling layers.

We denote an intermediate feature map from a specific layer as $H \in \mathbb{R}^{m \times dim}$ where m is the spatial dimensions dim is the dimensionality of each feature vector in the feature map and we denote a feature vector of j-th spatial dimension as H_j. Each attribute vector y_i is first projected to d-dimensional vector $y'_i = Uy_i$ with a projection matrix U. Then the attention coefficient α_{ji} for

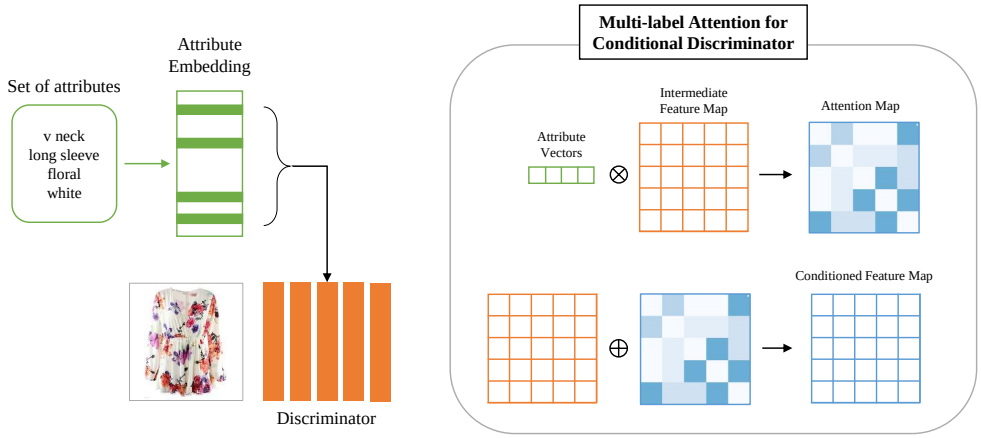


Figure 3.2 Illustration of Multi-label attention-based conditional discriminator of ADGAN

each feature vector H_j is computed by applying the softmax function to inner products between the feature vector H_j and condition attribute vectors y_i' as follows:

$$\alpha_{ji} = \frac{\exp(H_j \cdot y_i')}{\sum_{k=1}^{K} \exp(H_j \cdot y_k')}, \tag{3.1}$$

where K is the number of attributes in a condition attribute set.

This coefficients learn how much each attribute is related to synthesizing an image according to each spatial region. For example, for the upper region features of an product image, the coefficient of attributes related to *neckline* would be high, while for the left and right region *sleeve length* attributes would be tightly involved. Finally the context vector of each j-th spatial feature is calculated as weighted sum of projected attribute vectors as below:

$$c_j = \sum_{i=1}^{K} \alpha_{ji} y_i', \tag{3.2}$$

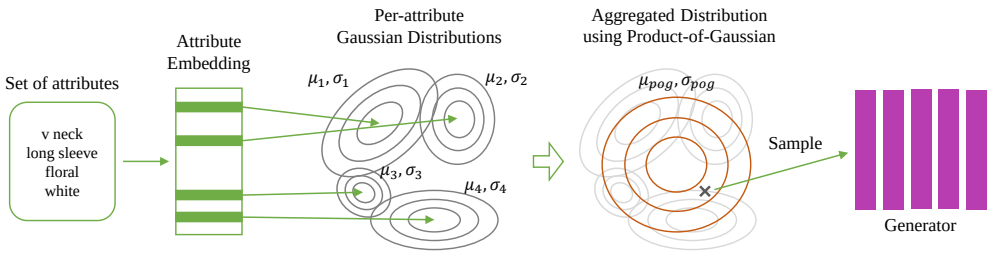


Figure 3.3 Illustration of Product-of-Gaussian based conditional prior sampling of ADGAN

and it is concatenated to the image feature H_j and the concatenated feature map is processed by next convolution layer. Detailed network structures are described at Table 7.1 and Table 7.2 in Appendix.

3.3.2 Product-of-Gaussian Condition Prior for Generator

Our aim is to generate as diverse images as possible while maintaining the consistency with given attribute combination. Vedantam et al. (Vedantam et al., 2018) define this concept as compositional abstraction hierarchy meaning conditionally generated samples should cover unobserved attributes. For example, if *neckline* and *color* is given as a condition, model should be capable of generate images with diverse *sleeve length* or *pattern*. To tackle this problem, we propose to use Product-of-Gaussian (Hinton, 2002) condition prior as a prior for GAN's generator.

As described in Figure 3.3, multiple independent Gaussian distributions are derived by mapping each attribute vector a_i to mean vector μ_i and diagonal covariance matrix σ_i. Then the prior distribution for attribute combination is obtained by multiplying all the conditional Gaussians. Since each independent

distribution is Gaussian, the product is itself Gaussian. The resulting Gaussian distribution is computed as follows:

$$\mu = \left(\sum_i \mu_i \sigma_i^{-1} \right) \left(\sum_i \sigma_i^{-1} \right)^{-1} \sigma = \left(\sum_i \sigma_i^{-1} \right)^{-1}, \tag{3.3}$$

and the conditional noise vector sampled from this Gaussian prior is mapped to image space by generator network.

Recently, Product-of-Gaussian is applied to Variational Auto-encoder based models in context of multimodal data generation (Vedantam et al., 2018; Wu and Goodman, 2018), but this is the first attempt to exploit Product-of-Gaussian to encode combination of attributes for GAN model to the best of our knowledge.

3.3.3 Visual-Semantic Embedding

Since the semantic meaning of attributes tend to be domain-specific, the attribute embedding also should be learned in a domain-specific way rather than naively exploiting word embeddings trained on general text corpus. So we build a visual-semantic embedding model to learn domain-specific attribute embeddings by associating the attributes to corresponding visual features via visual-semantic embedding learning.

For a pair of image x and a corresponding attribute set $y = \{y_1, y_2, ..., y_n\}$, let e_i represents an one hot vector of attribute y_i and $v = F(x)$ denotes an image feature vector of an image x produced by a convolutional neural network F and W is an attribute embedding matrix that we aim to learn. The aggregated feature vector a of given attribute set is computed by averaging the attribute feature vectors in the set:

Table 3.1 Number of images with different number of associated attributes in Polyvore dataset

# of attributes	1	2	3	4	5	6
# of images	20,734	8,521	2,222	409	50	9

$$a = \frac{1}{n} \sum_{i} a_i, \tag{3.4}$$

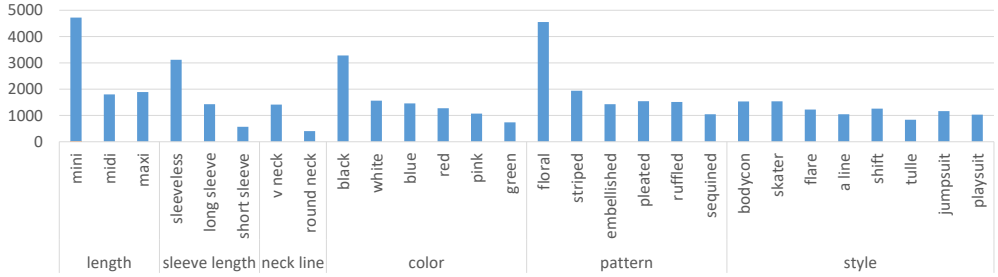
where $a_i = W \cdot e_i$ is the attribute embedding vector.

Our visual-semantic embedding model tempts to make paired image and attribute set features more similar than unpaired ones, so we define a similarity measure between image and attributes as cosine similarity, i.e., $d(v, a) = \frac{v \cdot a}{||v||_2 \cdot ||a||_2}$). Following Kiros et al. (2014), we train the image feature extractor F and the attribute embedding W by minimizing the bi-directional ranking loss as follows:

$$\begin{aligned} &\sum_{v,k} \max(0, m - d(v, a) + d(v, a_k)) + \\ &\sum_{a,k} \max(0, m - d(a, v) + d(a, v_k)), \end{aligned} \tag{3.5}$$

where a_k denotes non-matching attribute vectors for image feature v, v_k denotes non-matching image feature vectors for attribute set y and m is some margin. We use this pre-trained attribute embedding matrix W to extract attribute vectors for ADGAN's discriminator and generator.

Figure 3.4 Number of training images for each attribute in Polyvore dataset



3.4 Experiment

Dataset To validate proposed model, we collected 31,945 dress images and associated product title from a popular fashion site *polyvore.com*. To obtain attributes associated with each product image, we tokenize the title text into words and choose 6 criteria for fashion attributes by analyzing the tokenized words, which are *length, sleeve length, neckline, color, pattern, style*. Then we select total 28 attributes which of each belongs to one of 6 criteria. Figure 3.4 shows the image frequency of 28 attributes in each group and Table 3.1 shows the number of images those with different number of associated attributes. 28,750 image-attribute set pairs are used for training the model and the rest are used to evaluate it.

Evaluation Setup It is difficult to evaluate the generative models like GANs because there is no concrete and objective criterion to determine the quality of generated samples. Inception Score (Salimans et al., 2016) is widely used to quantify the image quality, which measures the KL divergence between the marginal distribution and class conditional distribution computed with pre-trained Inception network. Although Inception Score is widely accepted as

evaluation metric, fashion images are highly domain-specific, so it is difficult to evaluate them with Inception network trained on general images (e.g., ImageNet). In addition, Inception score cannot measure the alignment between generated images and the given conditioning attributes.

Instead we train oracle attribute classifier and use it to judge generated images whether they are well-aligned with conditioning attributes. Resnet(He et al., 2016a) is used to map product images to 28 dimensional vector and binary cross entropy loss is computed for each attribute dimension to be optimized.

Evaluation Metric We verify the quality of generated images, under three criteria which are correctness, coverage and degree of mode collapse.

Correctness measures how much the generated images are consistent with the given attributes. Let $C = \{a_1, ..., a_n\}$ denotes the condition attribute set and $X_C = \{x_1, ..., x_m\}$ denotes the set of generated images with given C. We compute correctness using oracle attribute classifier as follows,

$$correctness(X_C, C) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \hat{y}_{a_j}(x_i) \tag{3.6}$$

where $\hat{y}_a(x)$ is binary prediction on attribute a by oracle classifier for input image x. Well-conditioned generated images get higher correctness score.

Although oracle classifier is able to quantify the correctness, still it is not a perfect judge. Therefore, we also conduct human evaluation. We randomly select 300 attribute sets from the validation set. ADGAN and StackGAN (baseline) generates a single synthetic image conditioned on each attribute set. Then we assemble the images according to each attribute and ask 10 users (not including any of the authors) to answer whether the images are correspond to given attribute or not. The average correctness is calculated and we denote this score as $correctness_{Human}$.

Besides, it is meaningful to measure how diversely images are generated while maintaining the consistency with condition attributes. Vedantam et al. (2018) defines coverage metric to quantify the diversity of generated images. Main idea is to check that the attributes that are not specified vary across the generated images by comparing the true distribution p_U and the empirical distribution q_U over unobserved attributes $U = A \backslash C$ where A is the set of all the attributes. We use Jensen-Shannon divergence to compare the two distributions, so the coverage is defined as follows,

$$coverage(X_C, C) = 1 - JS(p_U, q_U(X_C)) \tag{3.7}$$

We get p_U by normalizing the attribute frequency of training data for attributes in U. $q_U(X_C)$ is obtained similarly by aggregating the number of attributes those are classified to be true for all the images in X_C by oracle classifier.

3.4.1 Quantitative Result

To validate our proposed ADGAN, we compare our results with Stage-I network of StackGAN (Zhang et al., 2017). While original StackGAN uses LSTM encoder to obtain conditioning sentence vector, instead of sentence vector, we average the attribute vectors in the set to make a condition vector c. StackGAN's generator receives this condition vector c and map it to gaussian mean μ and covariance vector σ to obtain a gaussian distribution. After c' is sampled from $N(\mu, \sigma)$, generator uses concatenated vector of c' and noise vector z as generators input prior. Discriminator receives the condition vector c and concatenate the condition vectors to output features from 3rd convolutional layer.

ADGAN has the same up-sample, down-sample network structures with

Method	Correct	$\text{Correct}_{\text{Human}}$	Coverage
Comparison to the baseline.			
StackGAN (Zhang et al., 2017)	0.495	0.475	0.919
ADGAN (Ours)	**0.698**	**0.624**	0.873
Ablation result for proposed conditional modules.			
ADGAN (w.o. attention)	0.530	-	**0.930**
ADGAN (w.o. PoG)	0.602	-	0.887
Ablation result for attention layers.			
ADGAN (attention on 1st)	0.667	-	0.906
ADGAN (attention on 2nd)	0.670	-	0.898
ADGAN (attention on 3rd)	**0.698**	**0.624**	0.873

Table 3.2 Quantitative comparison result

StackGAN except for ADGAN's two components, attentional discriminator and PoG prior. It is worth noting that our two components are orthogonal to the GAN's generator and discriminator design for attribute-to-image generation. We also validate ADGAN model without each component to test how much each component contributes to the performance. In addition, ADGAN variants with attention on different layer are also evaluated.

Table 3.2 shows quantitative results of ADGAN and baseline methods. As shown in results, ADGAN with attention on 3rd layer achieves the best correctness score, meaning that ADGAN generates more consistent images with given attributes than StackGAN. In the perspective of correctness, attentional discriminator elevates the performance more than PoG prior on generator. ADGAN also outperforms StackGAN with respect to correctness measured by

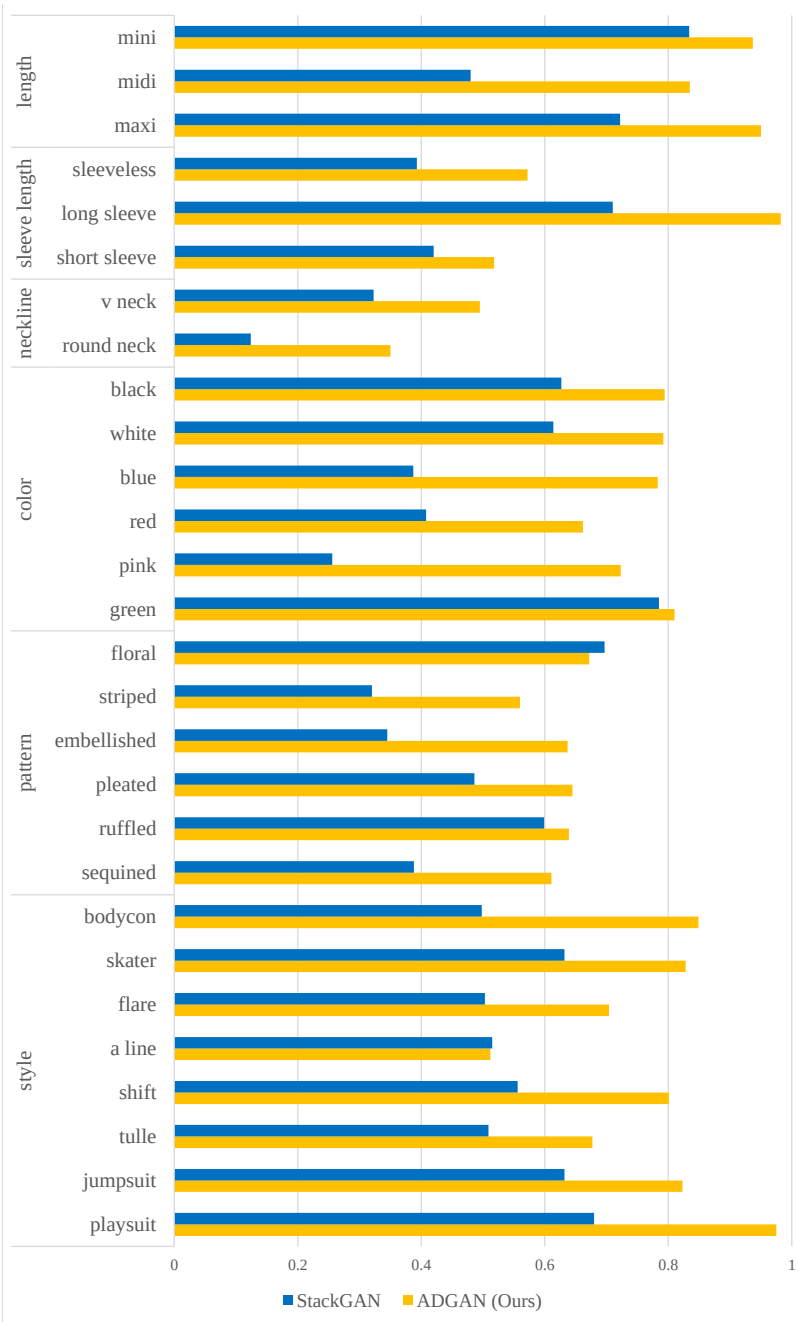


Figure 3.5 Comparison of correctness per attribute

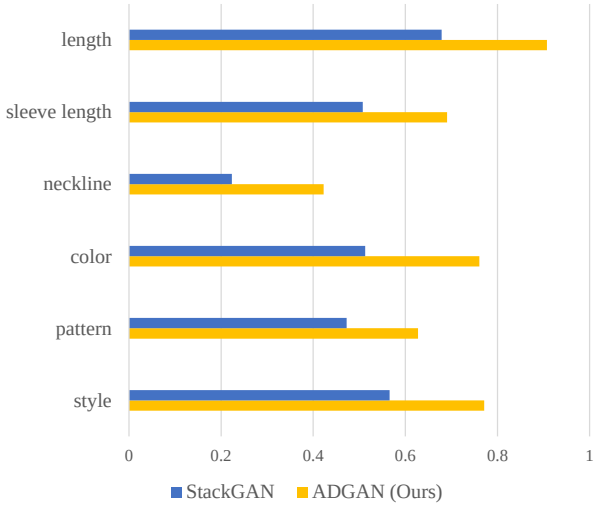


Figure 3.6 Comparison of correctness per attribute group

human evaluators. It is quite notable that the correctness score measured by oracle reaches the correctness measured by human users.

A trade-off between correctness and coverage score is observed, since the attributes have dependency to each other. For example, if an image is generated with attribute *skater dress*, it is a *mini* dress with a high probability. This kind of dependency between attributes makes the empirical distribution move away from the true distribution. Despite of such trade-off, we discovered that the Product-of-Gaussian prior increases the generalization capability. As demonstrated in Table 3.2, ADGAN with PoG prior generates more diverse images than ADGAN without PoG prior. In Table 3.2, ADGAN (w.o. attention) performs better than StackGAN and ADGAN (attention on 1st) performs better than ADGAN (w.o. PoG) with respect to coverage metric. Worth mentioning that the deeper layer on which the attention-plugged, the more extensive the effects of trade-off, so cautious selection of attention layer is needed.

Figure 3.5 presents the correctness score per attribute. It's not easy to state

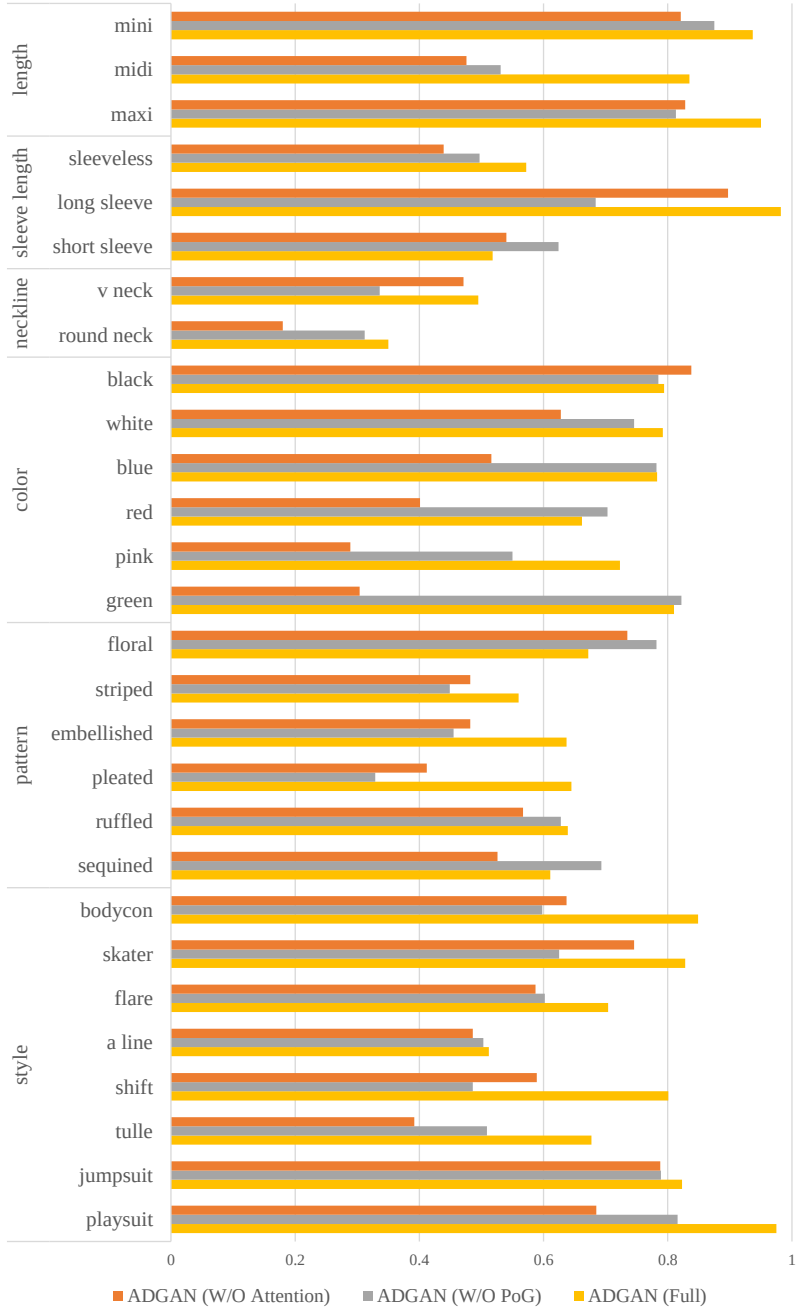


Figure 3.7 Ablation result of correctness per attribute

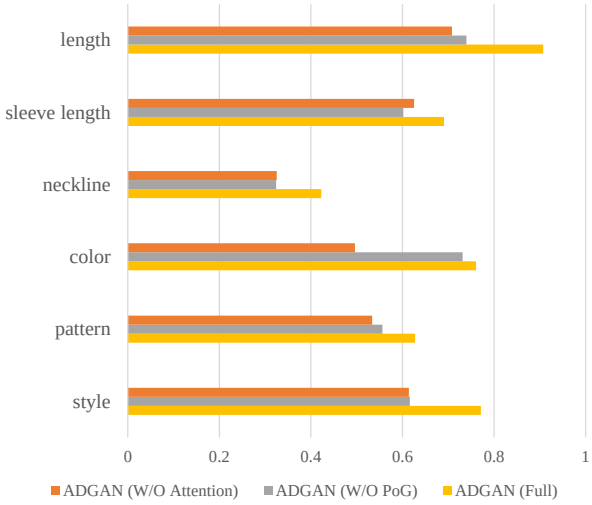


Figure 3.8 Ablation result of correctness per attribute group

clearly that one model outperforms the other for certain attribute or attribute group because the results fluctuate over attributes. As mentioned above, overall, ADGAN performs well. Especially, the attentional discriminator and PoG prior create synergy when generating images with *style* related attributes, because *style* is related to not only local patterns but also to the compositional features. For instance, most of *skater* dresses are *pleated mini* dresses. For attributes related to locally repeated patterns, like *colors* and *floral* patterns, the models with attention work well.

3.4.2 Qualitative Result

Besides quantitative evaluation, we also qualitatively examine the samples generated by the models. Figure 3.9 demonstrates the samples from StackGAN (top) and ADGAN (bottom) with same *color* attribute on each row. Both model generate images well-aligned with *black* and *white*, because those colors appear frequently in training set. However, in case of rare attributes (e.g.,

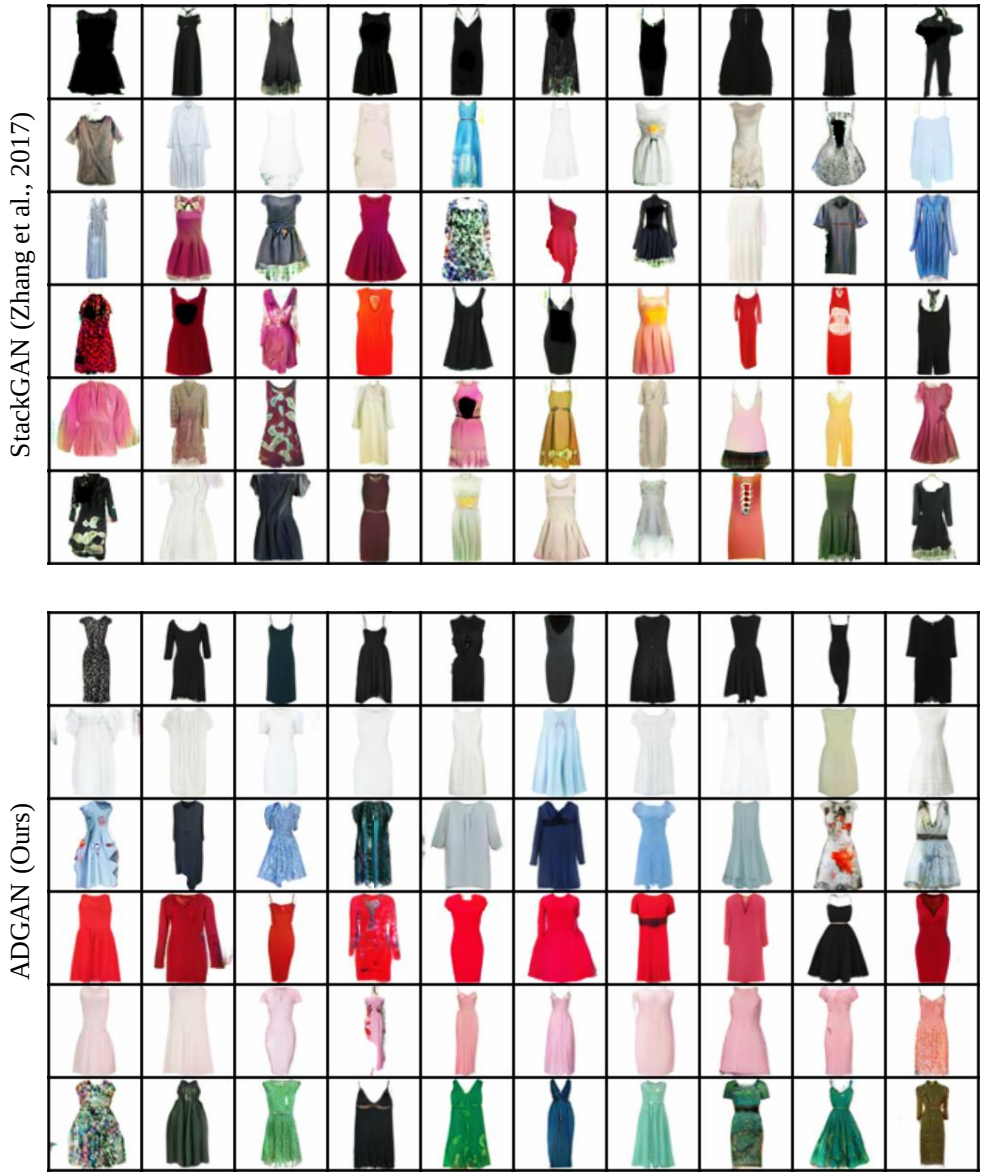


Figure 3.9 Qualitative comparison to the baseline

Figure 3.10 Images generated by ADGAN

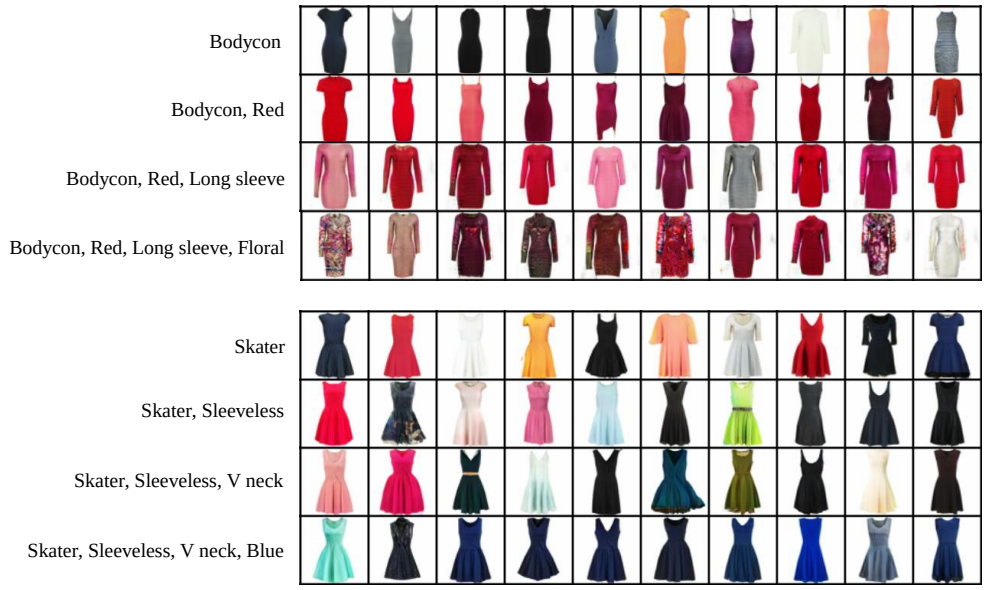


Figure 3.11 Images generated by ADGAN with a set of attributes

pink and green), StackGAN's results fail to maintain consistency with given condition while ADGAN produces coherent results. Figure 3.10 shows samples generated by ADGAN. Each column corresponds to each attribute in Figure 3.4 with same order. The result shows that the proposed ADGAN can generate images that are consistent to the given attribute condition.

One of the important features in multi attributes-to-image generation is a compositional abstraction ability. A compositional abstraction means that we can generate images of concepts at different abstraction level with different subsets of attributes. Figure 3.11 exhibits a sample generation from ADGAN showing such compositional abstraction ability. Top images are results produced by ADGAN with different attribute combinations from (*bodycon*) to (*bodycon*, *red*, *longsleeve*, *floral*) adding attributes one by one. The image set in the first row contains *bodycon* dresses those vary along different colors and shapes.

Second row consists of *bodycon* dresses with *red* color which are generated under more specified abstraction level and so on.

3.5 Chapter Summary

In this chapter, we explore a way to improve conditional generative model for complex multi-attribute conditions. To this end, we propose two modules for conditional discriminator and conditional generator. We leverage attention mechanism to the discriminator to better discriminate real and generated images based on given multiple attributes. We also propose a product-of-Gaussian based conditional prior for the generator. Both building blocks significantly boost the controllability on image generation, which is shown by experiment conducted on fashion images with complex attribute conditions.

Chapter 4

Multi-scale Contrastive Learning for Complex Scene Generation

4.1 Motivation

In recent years, generative adversarial networks (GAN) (Goodfellow et al., 2014) have achieved significant improvements due to extensive studies on network structures (Radford et al., 2015; Zhang et al., 2019; Brock et al., 2019; Karras et al., 2019, 2020b; Schonfeld et al., 2020), objective functions (Mao et al., 2017; Arjovsky et al., 2017; Lim and Ye, 2017a), and regularization techniques (Gulrajani et al., 2017; Miyato et al., 2018; Mescheder et al., 2018b). Now GAN models can produce high-quality images that are almost indistinguishable from real ones, showing impressive results in the wide range of object classes including human faces (Karras et al., 2019), animals (Brock et al., 2019; Schonfeld et al., 2020), and cars (Karras et al., 2020b). Despite these successes, when it comes to more complex images such as scenes with multiple objects, they easily fail to achieve the same level of realism as in single object images (Casanova et al.,

2020; Gadde et al., 2021).

In single object images, there is a common layout of each component, allowing it easier for the discriminator to supervise where and how each component should be synthesized to result in a realistic image. For instance, each component of dog's face, e.g., eyes, nose, and mouth, may vary in shapes and proportions, but remain in a common layout that forms the face. On the other hand, natural scene images exhibit much more diverse and complex distributions as they include a collection of objects in various sizes, shapes, and spatial locations (Casanova et al., 2020; Sylvain et al., 2021; Hua et al., 2021). Therefore, it is much harder for the discriminator to learn multi-layered differences between real and fake images from local semantic structures, such as objects, to overall scene layouts (Schönfeld et al., 2021; Gadde et al., 2021). As a result, even state-of-the-art GAN models produce unsatisfactory results of limited distribution coverage and low synthesis quality with messy layouts and incomplete internal objects.

In this chapter, we propose a method to improve discriminative ability on such complex scenes through a self-supervised pretext task assigned to the discriminator. Self-supervised representation learning has been extensively studied in recent years and shown to yield beneficial representations for various downstream tasks (Chen et al., 2020; He et al., 2020; Grill et al., 2020). The progress continues to generative models and recent studies have shown GAN models also can be improved by leveraging various self-supervised pretext tasks such as rotation prediction (Chen et al., 2019; Tran et al., 2019; Hou et al., 2021), consistency regularization (Zhang et al., 2020; Zhao et al., 2021) and contrastive learning (Zhao et al., 2020b; Jeong and Shin, 2021; Yang et al., 2021). While successful, existing studies mainly focus on enhancing image-level global representations especially for single-object images, thus the improvement tend to be

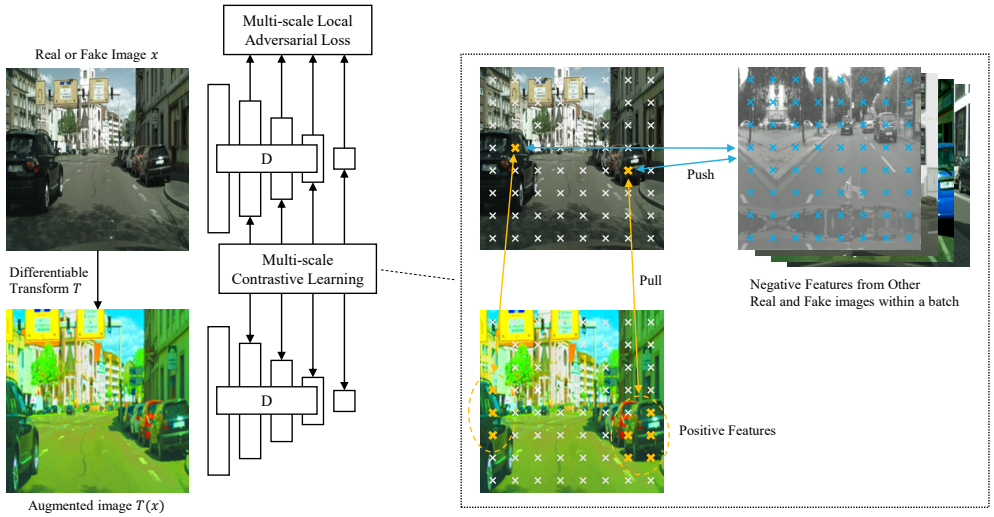


Figure 4.1 Illustration of proposed MsConD

limited for more complex data distributions, such as scene images containing various local objects.

To better model complex local semantic structures in the scene images, we propose to enhance local representations as well as the global representation with auxiliary pretext tasks locally defined and at multiple scales. To this end, we design a multi-scale discriminator having multi-level branches where each branch processes local patches of different sizes. Branch at each scale produces per-pixel auxiliary representations as well as per-pixel discriminator logits. These auxiliary representations are used to perform pixel-level contrastive learning to enhance per-pixel classification task. Both tasks are defined for each scale level and jointly optimized across all scales, thereby the discriminator could improve local-to-global discriminative ability to better model local structures in complex scenes at various scales. Figure 4.1 shows the overall architecture of proposed method.

We evaluate our method on several challenging scene image datasets with metrics for both scene-level and object-level synthesis quality. Compared to recent state-of-the-art GAN models, our method consistently achieves better results in terms of visual quality and diversity. In particular, our method significantly improves synthesis quality of individual objects in the scene, demonstrating that multi-scale representation learning effectively enhances the adversarial feedback to better model local semantic structures.

4.2 Related Work

Discriminator Design for GAN. Discriminator's ability to distinguish between real and fake images plays a critical role in GAN training, since the generator entirely relies on the feedback signal passed from the discriminator. Such ability has been significantly improved with the advances in discriminator architectures, from multi-layer perceptrons (Goodfellow et al., 2014) to convolutional networks (Radford et al., 2015; Karras et al., 2018), residual networks (Miyato et al., 2018; Karras et al., 2019), and self-attention based models (Zhang et al., 2019; Brock et al., 2019; Yu et al., 2021). However, even state-of-the-art models still struggle in modeling complex scenes, since they rely solely on global discriminator feedback therefore missing high frequency details. To alleviate the problem, we redesign the discriminator to utilize local feedback on multiple scales.

Local discriminator feedback has been used in various conditional image generation tasks (Zhu et al., 2017a; Huang et al., 2018; Park et al., 2019; Demir and Unal, 2018; Yu et al., 2019) in the form of PatchGAN discriminator (Isola et al., 2017). To cover multiple scales, Wang et al. (Wang et al., 2018) propose to use multiple PatchGAN discriminators to process each image interpolated at

different resolutions. These architectures have been helpful for modeling high frequency patterns, but they rely on explicit conditions such as segmentation maps or input images, to model global layouts. In contrast, our method allows to model local-to-global structures by utilizing multi-scale feedback which emerges from natural hierarchy inherent in the pyramidal features of backbone network. Recently proposed ProjectedGAN (Sauer et al., 2021) has also verified the usefulness of multi-scale features, but they focus on mixing multiple levels of pretrained features rather than utilizing local feedback.

Self-supervised Learning for GAN. Self-supervised learning has been recognized as one of the most influential methodologies in recent years as it can learn informative representations from a large amount of unlabeled data. Recent studies have shown that GAN training can also benefit from various self-supervised pretext tasks. A group of works (Chen et al., 2019; Tran et al., 2019; Hou et al., 2021) have shown that the rotation prediction task prevents catastrophic forgetting in GAN and leads to better results. Consistency regularization (Zhang et al., 2020; Zhao et al., 2021) stabilizes GAN training by imposing consistency of discriminator output between a clean image and its augmented version. More recently, several studies have explored the use of the instance discrimination task (Wu et al., 2018; He et al., 2020; Chen et al., 2020) as an auxiliary task to further enhance the discriminator (Zhao et al., 2020b; Jeong and Shin, 2021; Yang et al., 2021). The self-supervised pretext tasks generally involve various image transformation functions to acquire different views of an image. In GAN training, differentiable image transformations (Karras et al., 2020a; Zhao et al., 2020a) applied on both real and fake images have shown to stabilize the training in limited data regimes and improve the data efficiency. Our work relies on previous findings on improved GAN training

with self-supervised pretext tasks. However, while all previous studies focus on enhancing global representation space by integrating image-level tasks, in this chapter, we seek to enhance region-level representations to improve discriminative ability on local features.

Dense Representation Learning. Recent studies on self-supervised representation learning mainly focus on image-level representations for object-centric images, i.e., ImageNet (Deng et al., 2009). Despite their success, the image-level global representations are often sub-optimal for general vision tasks defined on complex scenes, as globally pooled representations lose spatial information of local objects. Therefore, more recent works attempt to learn pixel-level (Pinheiro et al., 2020; Xie et al., 2021c; Wang et al., 2021b) or region-level (Roh et al., 2021; Xiao et al., 2021; Xie et al., 2021a) representations and have achieved meaningful improvements in dense prediction downstream tasks such as object detection and instance segmentation. We repurpose the dense representation learning as a mean to aid the real-fake discrimination on multiple scales, thereby validate its efficacy on improving synthesis quality of local objects in complex scenes.

4.3 Method

In this section, we describe the proposed method, namely, Multi-scale Contrastive Discriminator (*MsConD*) in detail. We describe the improved discriminator architecture in Section 4.3.1, followed by multi-scale pixel-level contrastive learning that further enhances the discriminator in Section 4.3.2 and finally the full objective function which optimizes the entire network in Section 4.3.3.

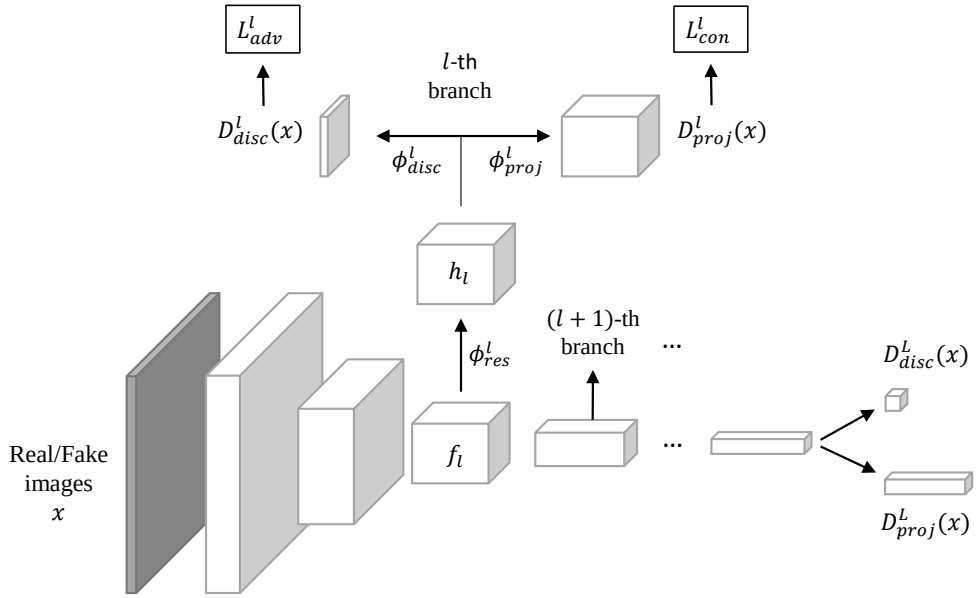


Figure 4.2 Illustration of proposed discriminator architecture

4.3.1 Multi-scale Discriminator with Multi-level Branches

In unconditional image synthesis, a discriminator is typically equipped with several sub-sampling layers that progressively downsample the input high-resolution images into lower resolution features constructing pyramidal feature maps (Radford et al., 2015; Brock et al., 2019; Zhang et al., 2019; Karras et al., 2019). To enable discrimination of each local feature in the feature maps, we use branches for each scale l to translate the intermediate features into corresponding local outputs. Each branch consists of three components: residual blocks ϕ^l_{res}, a classification head ϕ^l_{disc}, and a projection head ϕ^l_{proj}. All components are implemented with 1×1 convolution layers to process each local feature individually. Figure 4.2 shows the proposed discriminator design.

Concretely, our discriminator D is composed of backbone network F and

per-scale branch networks $\phi^l = \{\phi^l_{res}, \phi^l_{disc}, \phi^l_{proj}\}$. Given an input image, the backbone network F produces multi-scale feature maps. We denote the feature map at scale level l as f_l. f_l is first transformed into h_l of the same shape by ϕ^l_{res} and then h^l is processed by two separate head networks, a real/fake classification head ϕ^l_{disc} and a projection head ϕ^l_{proj} to produce two outputs U^l and V^l.

$$h_l = \phi^l_{res}(f_l) \in \mathbb{R}^{H_l \times W_l \times C_h} \tag{4.1}$$

$$U_l = \phi^l_{disc}(h_l) \in \mathbb{R}^{H_l \times W_l \times 1} \tag{4.2}$$

$$V_l = \phi^l_{proj}(h_l) \in \mathbb{R}^{H_l \times W_l \times C_p}, \tag{4.3}$$

where C_p is number of channels of the projection output.

We denote the classification head output U_l and the projection output V_l for an input image x as $D^l_{disc}(x)$ and $D^l_{proj}(x)$, respectively. $D^l_{disc}(x)$ is used to compute per-pixel adversarial loss at l-th scale while $D^l_{proj}(x)$ is used to perform pixel-level contrastive learning which will be described in the following section. The adversarial loss at l-th scale is computed by averaging all per-pixel adversarial losses as follows:

$$\begin{aligned}\mathcal{L}^l_{adv}(G, D) = \mathbb{E}_x \Bigg[\frac{1}{H_l W_l} \sum_{i,j} \log \Big[D^l_{disc}(x) \Big]_{i,j} \Bigg] \\ + \mathbb{E}_z \Bigg[\frac{1}{H_l W_l} \sum_{i,j} \log \Big(1 - \Big[D^l_{disc}(G(z)) \Big]_{i,j} \Big) \Bigg],\end{aligned} \tag{4.4}$$

where $\big[D^l_{disc}(x)\big]_{i,j}$ refers to the classification output at pixel (i, j). As shown in Figure 4.2, the global representation at the top of the backbone network is likewise mapped to the global discriminator output and the global projection output, which are used to compute the adversarial and contrastive losses,

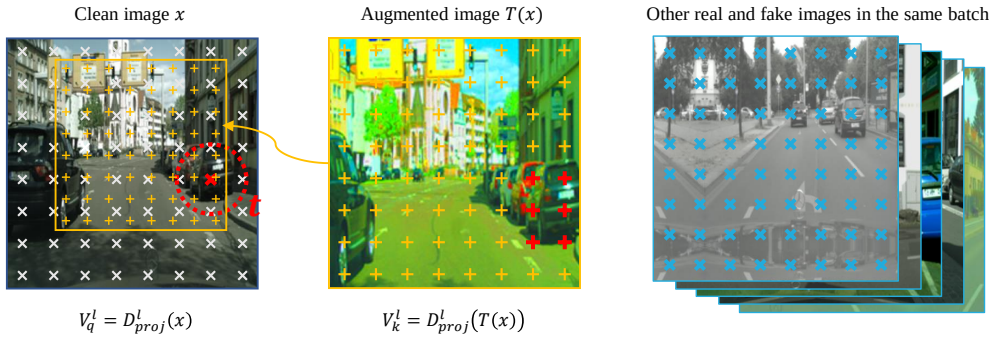


Figure 4.3 Illustration of spatially consistent pixel-level contrastive learning in MsConD

respectively. See Table 7.3 in Appendix for more detailed network architecture.

4.3.2 Multi-scale Contrastive Learning for GAN

The redesigned discriminator learns to differentiate between real and fake images based on local-to-global region-level decisions. To further enhance the discriminative ability, we propose to assign the discriminator an auxiliary self-supervised task designed to enrich the region-level representation on which each decision is performed.

Given a clean image x, its augmented view $T(x)$ is obtained by applying a differentiable transformation T. Then the respective projection outputs V_q^l and V_k^l at l-th scale are extracted through the projection branch:

$$V_q^l = D_{proj}^l(x) \in \mathbb{R}^{H_l W_l \times C_p} \tag{4.5}$$

$$V_k^l = D_{proj}^l\left(T(x)\right) \in \mathbb{R}^{H_l W_l \times C_p}. \tag{4.6}$$

Instance discrimination task (Wu et al., 2018; He et al., 2020; Chen et al., 2020) is a widely adopted pretext task in self-supervised representation learning.

Typically, it conducts training by contrasting the positive views of an instance from the negative views which are irrelevant to the instance. In image-level instance discrimination task, the positive features can be easily obtained by simply applying random transformations to an image. However, our objective is to learn local representations to support real-fake decision on individual local features, thereby an instance for the task no longer represents the whole image but local regions of an image. In this case, the positive features should be cautiously identified to ensure sufficient overlap between the regions represented by the features. Otherwise, it can interfere with representation learning by associating areas that are completely unrelated to each other in the image.

In this chapter, we identify two feature vectors from V_q^l and V_k^l as a positive pair if they are close enough to contain the same region in the image (Xie et al., 2021c). The spatial closeness is measured by the Euclidean distance between the coordinates of two feature vectors in the image space. Figure 4.3 shows an example. Concretely, we warp the pixels in V_k^l into the clean image space to obtain the reference coordinates and compute all-pair Euclidean distances between the coordinates of feature vectors in the two feature maps V_q^l and V_k^l. For each feature vector $v_q \in \mathbb{R}^{C_p}$ in V_q^l, we define the positive feature set from V_k^l as follows:

$$pos(v_q) = \{v_k \in V_k^l : dist(v_q, v_k) < t\}, \tag{4.7}$$

where $dist(v_q, v_k)$ denotes the Euclidean distance between the coordinates of feature vectors v_q and v_k in the clean image space, and t is predefined distance threshold.

On the other hand, we construct the negative feature set $neg(v_q)$ with the same level features from other images in the same mini-batch. It is worth noting

that we use both real and fake images for negative features in order to construct larger negative set. We empirically observed that this leads to a slight performance improvement. With positive and negative feature sets, the contrastive loss at l-th layer can be formulated as:

$$\mathcal{L}^l_{con}(x, T(x)) = \sum_{v_q \in V^l_q} -\log \frac{\sum\limits_{v_k \in pos(v_q)} e^{v_q \cdot v_k/\tau}}{\sum\limits_{v_k \in pos(v_q)} e^{v_q \cdot v_k/\tau} + \sum\limits_{v_k \in neg(v_q)} e^{v_q \cdot v_k/\tau}}, \tag{4.8}$$

where τ is a temperature hyper-parameter which is set to 0.3. We normalize the feature vector v_q and v_k before computing the contrastive loss thus the dot product between them assesses the cosine similarity between the vectors.

We demand the discriminator to solve the same task for fake images $G(z)$ and their augmented views $T(G(z))$. In this way, the discriminator can learn from infinite samples generated by the model beyond the limited amount real images (Yang et al., 2021). Finally, the contrastive loss at l-th scale is computed using contrastive losses applied on both real and fake sample as follows:

$$\mathcal{L}^l_{con}(D) = \mathbb{E}_x \left[\mathcal{L}^l_{con}(x, T(x)) \right] + \mathbb{E}_z \left[\mathcal{L}^l_{con}\left(G(z), T(G(z))\right) \right]. \tag{4.9}$$

4.3.3 Full Objective

The total loss for MsConD is calculated using adversarial loss and contrastive loss summed on all scales as follows:

$$\mathcal{L}_{adv}(G, D) = \sum_{l} \mathcal{L}^{l}_{adv}(G, D) \tag{4.10}$$

$$\mathcal{L}_{con}(D) = \sum_{l} \mathcal{L}^{l}_{con}(D) \tag{4.11}$$

$$\min_{G} \max_{D} \mathcal{L}_{adv}(G, D) - \lambda \mathcal{L}_{con}(D), \tag{4.12}$$

where λ controls the strength of contrastive loss. We found that $\lambda = 0.2$ gives desirable balance between the two loss terms, and we use this value for all experiments.

4.3.4 Implementation and Training

The MsConD is implemented upon the resnet-based discriminator of StyleGAN2 (Karras et al., 2020b). We adopt the training techniques used in StyleGAN2 including lazy R1 regularization and path length regularization. For augmentation T, we use differentiable transformations including pixel blitting, geometric and color transformations following StyleGAN2-ADA (Karras et al., 2020a). One notable difference is that StyleGAN2 computes the R1 regularization loss using the global discriminator output, whereas MsConD computes the R1 losses for each branch output and regularizes the network with the sum of the losses. We use Adam optimizer with batch size of 32, learning rate of 0.002, $\beta_1 = 0.0$ and $\beta_2 = 0.99$. All models including the baselines have been trained for the same number of training steps (10 million images).

4.4 Experiment

Datasets We evaluate the proposed method on three challenging scene image datasets. Cityscapes (Cordts et al., 2016) contains 25k images of street scenes

recorded from a driving car in 50 cities. LSUN (Yu et al., 2015) is a large collection of scene images covering wide range of indoor and outdoor scenes. Among them, we choose livingroom and kitchen dataset as benchmark datasets since they exhibit highly complex data distributions derived from diverse scene layouts with various objects. Livingroom and kitchen datasets contain 1.3 million and 2.2 million scene images, respectively. All images used in the experiments are resized to 256×256 resolution.

Evaluation metrics To quantitatively evaluate the synthesis quality, we use Frechet inception distance (Heusel et al., 2017a), Kernel inception distance (Bińkowski et al., 2018), Precision, and Recall (Kynkäänniemi et al., 2019). Following the previous works (Heusel et al., 2017b; Karras et al., 2020a), all metrics are calculated using 50,000 fake images and all training images.

Perceptual quality of scene images is largely determined by synthesis quality of individual objects within the scene. Since there is no object-level label in the evaluation dataset, we employ a pretrained object detector to identify objects depicted in both real and generated scenes. Then we calculate FID scores using the crops of detected objects for each object category. The object crops detected from 50,000 real images are used to obtain per-category real distributions. For a fair comparison, we calculate the FID using the same number of object crops from each model. We use YOLOR (Wang et al., 2021a) object detector trained on MS-COCO (Lin et al., 2014).

Comparison methods We use several recent competitive models as our baselines. UnetGAN (Schonfeld et al., 2020) and StyleGAN2 (Karras et al., 2020b) are utilized to compare different discriminator architectures. ADA (Karras et al., 2020a) uses differentiable data augmentations, while InsGen (Yang

	Cityscapes				Livingroom				Kitchen			
Method	FID↓	KID↓	Prec↑	Rec↑	FID↓	KID↓	Prec↑	Rec↑	FID↓	KID↓	Prec↑	Rec↑
UnetGAN (Schonfeld et al., 2020)	14.47	8.41	0.434	0.132	6.73	3.92	0.518	0.265	6.71	4.13	0.528	0.290
StyleGAN2 (Karras et al., 2020b)	8.04	5.27	0.539	0.260	4.64	2.22	0.512	0.268	5.10	2.58	0.530	0.305
ADA (Karras et al., 2020a)	5.03	1.86	0.604	0.221	4.95	2.34	0.507	0.267	6.47	3.62	0.484	0.272
InsGen (Yang et al., 2021)	4.21	1.64	0.583	0.349	4.17	2.09	0.556	0.318	5.76	2.57	0.535	0.312
ProjectedGAN (Sauer et al., 2021)	5.07	1.94	**0.620**	0.270	5.51	2.36	**0.571**	0.273	4.38	2.11	**0.587**	0.250
MsConD (Ours)	**2.63**	**0.99**	0.605	**0.485**	**2.73**	**1.14**	0.538	**0.431**	**2.88**	**0.97**	0.544	**0.429**

Table 4.1 Comparison result on Scene-level generation metrics

Cityscapes	car		person		traffic light		truck		bus	
	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓
StyleGAN2 (Karras et al., 2020b)	10.96	7.38	26.75	16.96	86.71	51.06	34.06	16.46	61.45	24.51
InsGen (Yang et al., 2021)	7.89	4.72	26.04	14.52	81.52	40.53	36.65	15.50	64.01	25.30
ProjectedGAN (Sauer et al., 2021)	20.12	11.12	32.59	30.41	96.51	32.78	57.14	12.39	76.50	37.46
MsConD (Ours)	**4.57**	**2.64**	**17.02**	**8.60**	**48.16**	**17.50**	**23.13**	**6.14**	**52.80**	**18.90**
Livingroom	couch		chair		potted plant		tv		vase	
	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓
StyleGAN2 (Karras et al., 2020b)	11.21	8.06	14.58	8.64	16.09	8.14	12.22	9.02	40.19	6.19
InsGen (Yang et al., 2021)	9.57	7.01	14.22	9.03	14.20	7.42	14.62	11.23	39.44	5.75
ProjectedGAN (Sauer et al., 2021)	8.60	4.68	21.77	10.18	22.16	12.03	12.76	7.12	42.87	6.44
MsConD (Ours)	**4.30**	**2.19**	**8.64**	**3.66**	**10.15**	**2.92**	**9.51**	**4.47**	**35.86**	**3.52**
Kitchen	oven		chair		microwave		potted plant		refrigerator	
	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓
StyleGAN2 (Karras et al., 2020b)	8.91	4.98	19.89	11.32	11.82	7.57	21.77	10.89	19.11	10.05
InsGen (Yang et al., 2021)	9.15	4.87	19.70	10.49	11.74	6.14	20.18	7.89	17.45	8.83
ProjectedGAN (Sauer et al., 2021)	13.78	6.01	21.74	9.42	20.06	12.29	25.95	8.24	21.55	13.60
MsConD (Ours)	**5.01**	**1.36**	**13.28**	**5.54**	**8.85**	**3.33**	**15.17**	**3.40**	**12.36**	**4.86**

Table 4.2 Object-level metrics for each object category

et al., 2021) applies image-level instance discrimination upon ADA. ProjectedGAN (Sauer et al., 2021) is a parallel state-of-the-art study using multiple discriminators to leverage multi-scale features from pretrained networks. We use officially released code base of baseline methods except for UnetGAN where we employ better backbone of StyleGAN2.

We use the same StyleGAN2 generator for all methods to fairly compare the discriminator ability except for ProjectedGAN where the lighter generator, i.e., FastGAN (Liu et al., 2021a) generator, has been reported to perform better. Since we observed that most methods are highly sensitive to the R1 penalty term (Mescheder et al., 2018b), we carefully explored the best performing R1 weights in the range of 1 to 50 for each method. For ADA and InsGen, we use the same set of image transformations consisting of pixel blitting, geometric and color transformations, which have shown the most stable results in the literature. For other hyper-parameters, we use the same values as originally proposed in each paper.

4.4.1 Comparison to State-of-the-Art

Scene-level Metrics. Table 4.1 shows the quantitative comparison result using standard GAN metrics. In terms of FID, our method outperforms all other baselines, achieving 37%, 35% and 33% relative improvements in each dataset compared to the best baseline methods. Our method achieves significantly improved recall across all datasets, demonstrating its capability to synthesize diverse scene images. Albeit ProjectedGAN achieves the highest precision, we empirically observed that it produces larger fraction of artifacts than other methods. This is also verified by its inferior object synthesis quality in Table 4.2. We speculate that the pretrained feature space learned on object-centric images, i.e., ImageNet, may not be best suited for learning more complex data

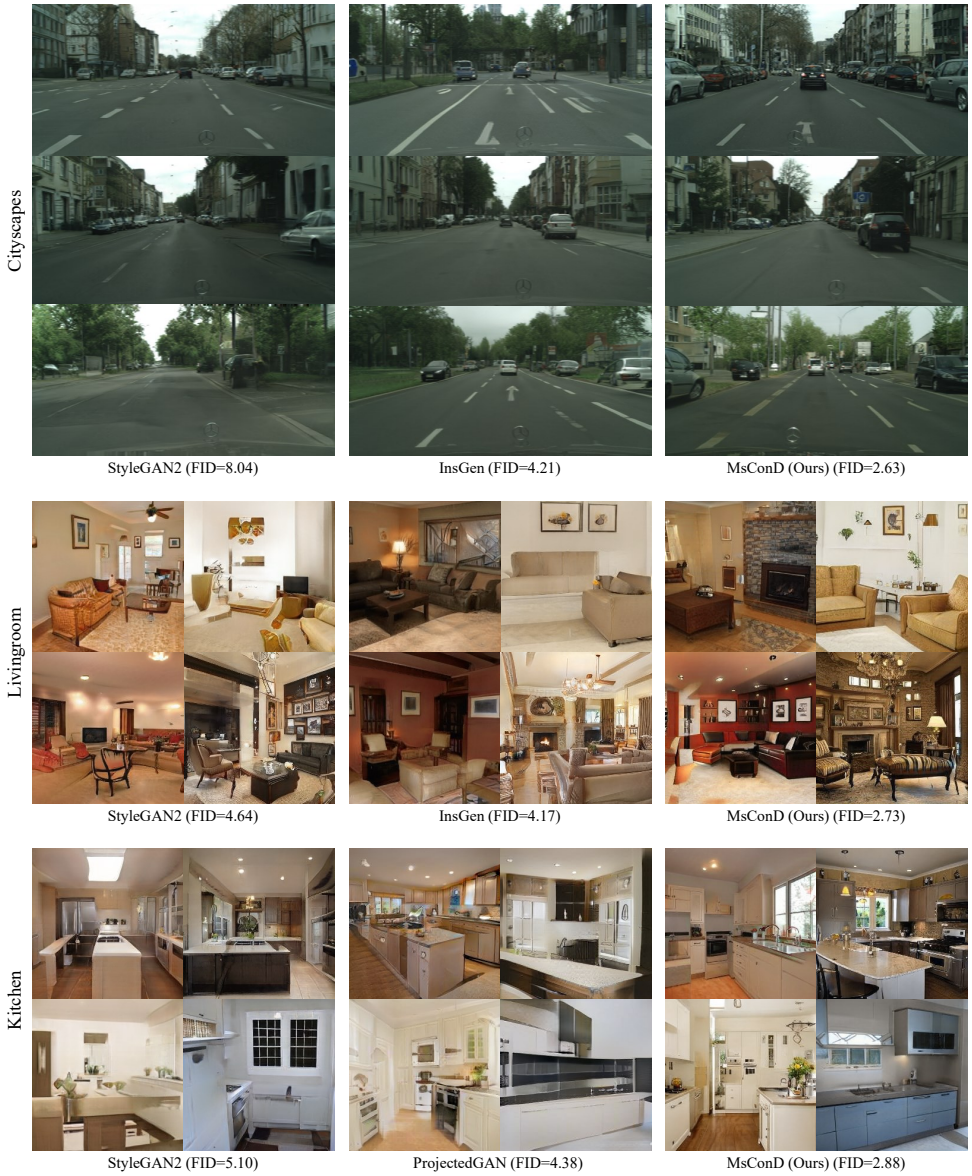


Figure 4.4 Comparison of generated samples

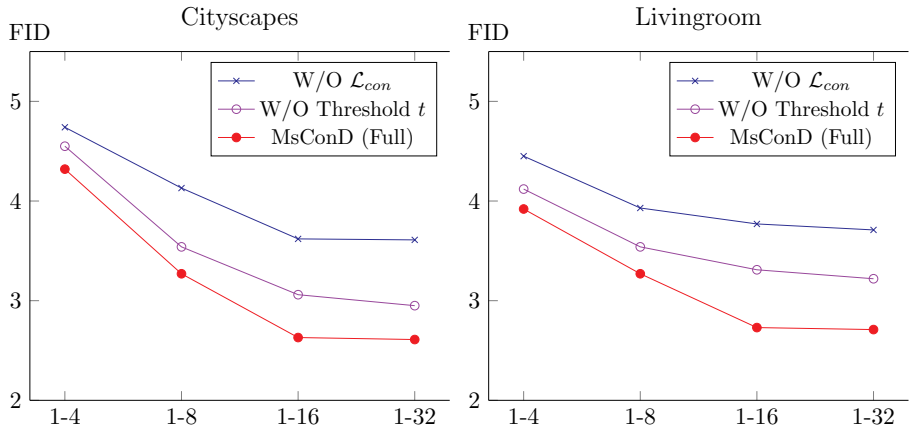
Figure 4.5 Samples Generated with MsConD

distributions.

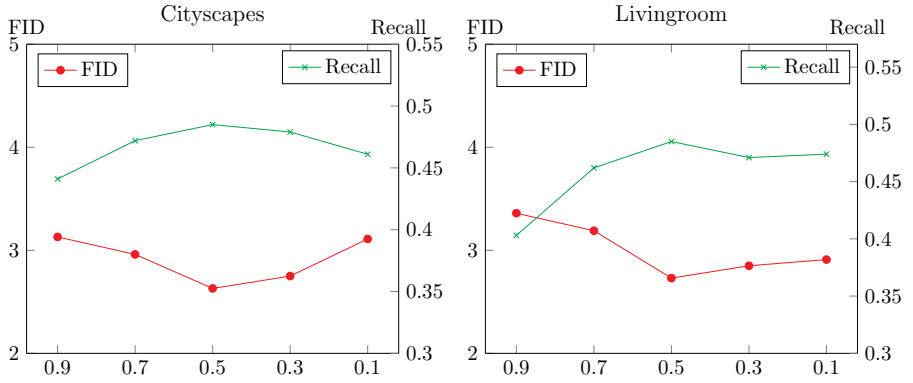
Object-level Metrics. To validate if our method improves the synthesis quality of individual objects in the scene, we measure FID, KID, and IS scores for top 5 most frequent object categories detected in each data domain. Table 4.2 shows the comparative result. In all object categories, our method achieves significantly improved metric scores over the baselines. These results validate that the proposed MsConD effectively incentivizes the generator to improve local details and produce more realistic objects in the scene images. Figure 4.4 provides visual comparison between samples generated by different methods. As shown, our method produces more realistic scene details with a well arranged layout over other methods. See Section 7.2 in Appendix for more result.

4.4.2 Ablation Study

In this section, we conduct an ablation study to investigate how each component of MsConD contributes to the generation performance. Figure 4.6 summarizes the ablation results. Figure 4.6 (a) shows scene-level FID when MsConD is trained with different scales of feature maps. We compare the model with/without multi-scale contrastive loss $\mathcal{L}_{con}$ to validate its efficacy. As shown in the result, in both cases, the generation performance increases as more feature maps are utilized, yet the performance is prominently boosted through multi-scale contrastive learning. We also report the result with contrastive learning but without distance threshold t to verify the effectiveness of our strategy for positive feature sampling. In this case, we use all pairs of local features from augmented images as positive samples without any spatial constraints. The result shows that the performance gain is far limited without the distance threshold, since semantically irrelevant local features impede the representation learning.



(a) Ablating each component of MsConD with varying scales.



(b) Effect of distance threshold t.

Figure 4.6 Quantitative Ablation Result

To further investigate the effect of distance threshold, we report FID and Recall with varying thresholds in Figure 4.6 (b). The performance deteriorates if t is too high or too low. When t is too low, only a narrow range of features are utilized as positive features, degrading the sample diversity. On the other hand, if the t is too high, irrelevant features could be treated as the positive features, and possibly hinder the learning.

Figure 4.7 shows quantitative ablation result for each object category. We

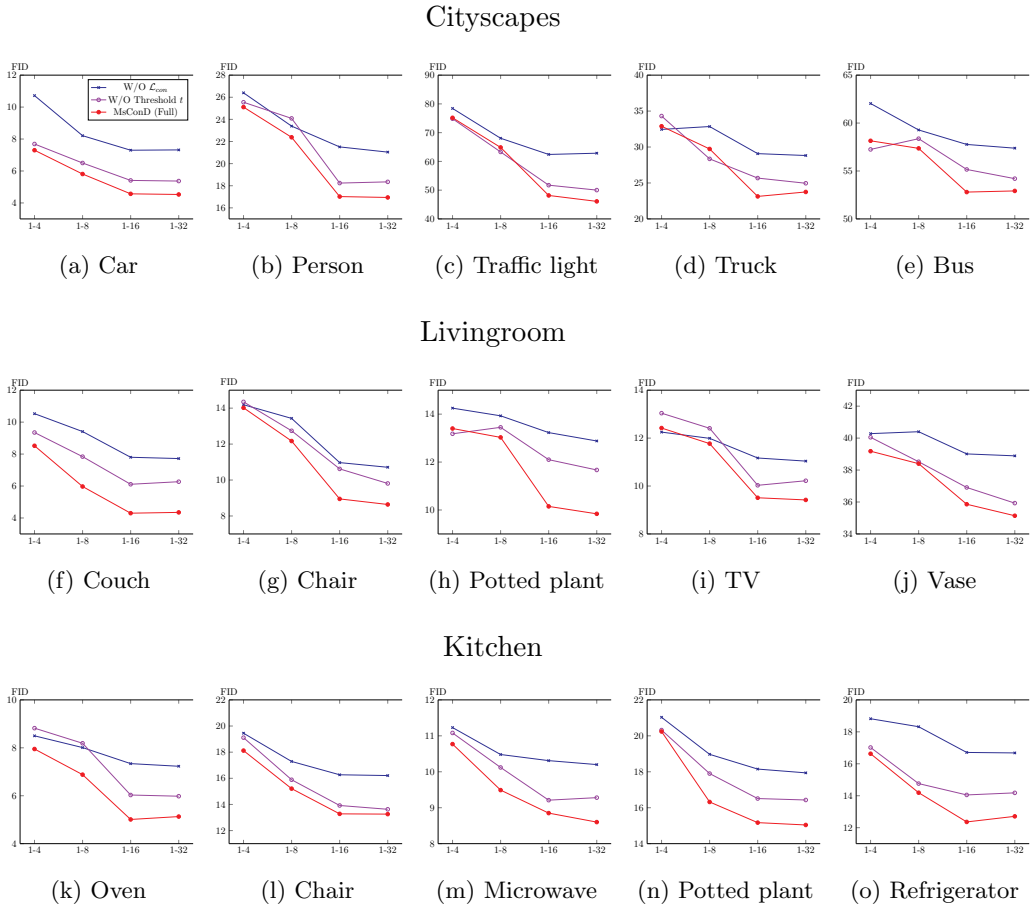


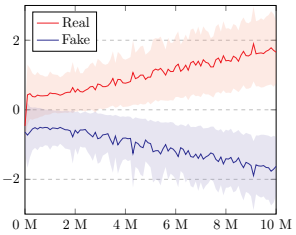
Figure 4.7 Quantitative ablation result for each object category

observe similar tendency as the scene-level ablation result in Figure 4.6. The generation performance increases as more feature maps from the backbone layers are utilized even when multi-scale contrastive learning is not applied. However, the performance is significantly improved as the contrastive learning is leveraged as an auxiliary task. The improvement has been consistent across various object categories validating the efficacy of MsConD in synthesizing local objects.

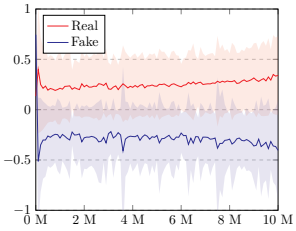
Figure 4.8 shows samples generated by MsConD trained under different

Figure 4.8 Qualitative ablation result on Livingroom dataset

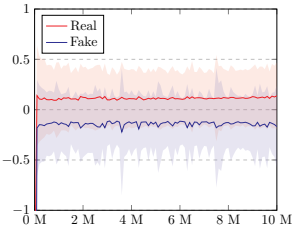
configurations. When the model is trained without multi-scale adversarial loss (W/O MS Adv.), local objects tend to be incomplete and discontinued as the generator is not provided with local feedback. On the other hand, when the model is trained only with multi-scale adversarial loss (W/O MS Con.), we observe repetitive patterns often appear in the generated images, which are known to be a common side-effect of PatchGAN discriminator. These artifacts are prominently mitigated in the results of MsConD, resulting in more realistic local objects.



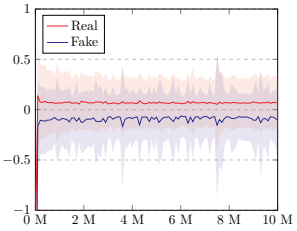
(a) StyleGAN2 $D(x)$



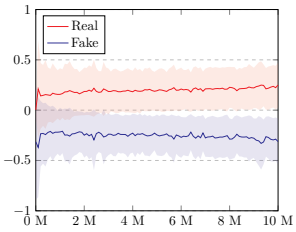
(b) MultiscaleD $D^{1}_{disc}(x)$



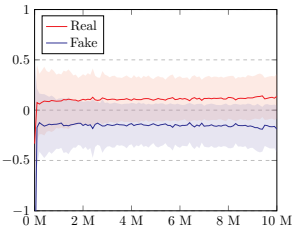
(c) MultiscaleD $D^{8}_{disc}(x)$



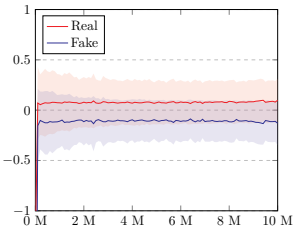
(d) MultiscaleD $D^{16}_{disc}(x)$



(e) MsConD $D^{1}_{disc}(x)$



(f) MsConD $D^{8}_{disc}(x)$



(g) MsConD $D^{16}_{disc}(x)$

Figure 4.9 Comparison of training progress on Cityscapes

4.4.3 Analysis on Training Dynamics

To further understand the training behavior of MsConD, we investigate the statistics of discriminator logits for real and fake images during the training process. Figure 4.9 shows the result on Cityscapes. For StyleGAN2, the logit distributions overlap during the initial training period and then gradually move away from each other. Therefore, as the training progresses, the discriminator becomes overly confident and fails to provide meaningful feedback to the generator, resulting in degraded synthesis quality. To mitigate the overfitting of discriminator, previous studies mainly focus on developing differentiable image augmentations (Karras et al., 2020a; Zhao et al., 2020a). Our findings indicate the problem could be substantially alleviated by utilizing multi-scale adversarial feedback.

Figure 4.9 (b-d) presents the results when local discriminator feedback is incorporated by our proposed discriminator. As shown, the logit distributions of real and fake samples remain within a close range for the entire training period, indicating that the discriminator can continuously provide informative feedback without overfitting. Meanwhile, we could observe that the fake logits for higher frequency part, i.e., $D^{16}_{disc}(x)$, tend to be unstable with large deviations. This instability stems from large structural variations of high frequency patterns in complex scenes. Figure 4.9 (e-g) shows that the auxiliary representation learning effectively stabilizes the feedback signal, in turn further improves the synthesis quality.

4.5 Chapter Summary

Despite recent advances of GANs, challenges still remain in modeling more complex data distributions. One of these challenges lies in learning complex

and diverse local structures, such as individual objects in scene images. To mitigate the difficulty, we redesign the discriminator to leverage local feedback from multi-scale features through multi-level branches. In addition, we propose to enrich the multi-scale representations through contrastive learning in order to further enhance the multi-scale GAN feedback. Experimental results show our method improves the local-to-global discriminative ability, thus effectively incentivizes the generator to synthesize diverse scene images with realistic details.

Chapter 5

Leveraging Pretrained Vision Models for Complex Scene Generation

5.1 Motivation

In this chapter, we explore another way to improve the complex scene generation. As mentioned in previous chapter, modeling the distribution of complex scenes is still challenging even for state-of-the-art GAN models and this difficulty lies in the high structural complexity of scene images. Therefore, during GAN training, the discriminator is under the heavy burden of learning complex structural differences between real and fake scene images to properly guide the generator.

In the perspective of recognition models, scene understanding tasks have been extensively studied over the past decades (Zou et al., 2019; Lateef and Ruichek, 2019). Now we have access to advanced models trained for various scene understanding tasks. A natural question follows: can we take advantage

of these pretrained scene recognition models to relieve the burden of the discriminator in GAN training and enhance its discriminative capability to better incentivize the generator. In this chapter, we explore a way to leverage pretrained scene understanding models to improve GAN models.

While transfer learning has been a ubiquitous process in facilitating downstream tasks, it is relatively under-explored for GAN models where most of GAN models are trained from scratch. There has been a recent study (Sauer et al., 2021) that utilized an ImageNet pretrained network as a feature extractor for the discriminator and improved the synthesis quality as well as the convergence speed. However, we observe that the performance gain is far limited for complex scene images since the utilized pretrained network is trained on object centric images, i.e., ImageNet (Deng et al., 2009). In this chapter, we explore the use of pretrained scene understanding models with the aim of improving complex scene generation. As the models trained on different scene understanding tasks contain distinct knowledge on visual scenes, we propose an feature level ensemble method to fully utilize the pretrained features from multiple models.

We validate the efficacy of proposed method on two challenging scene image datasets. Compared to recent GAN models, the proposed method consistently improves the synthesis quality of complex scenes verified by various generation metrics. In addition to scene-level metrics, the proposed method significantly improves object-level synthesis quality, demonstrating the pretrained expert models substantially aid the discriminator in recognizing meaningful semantic structures in the scene.

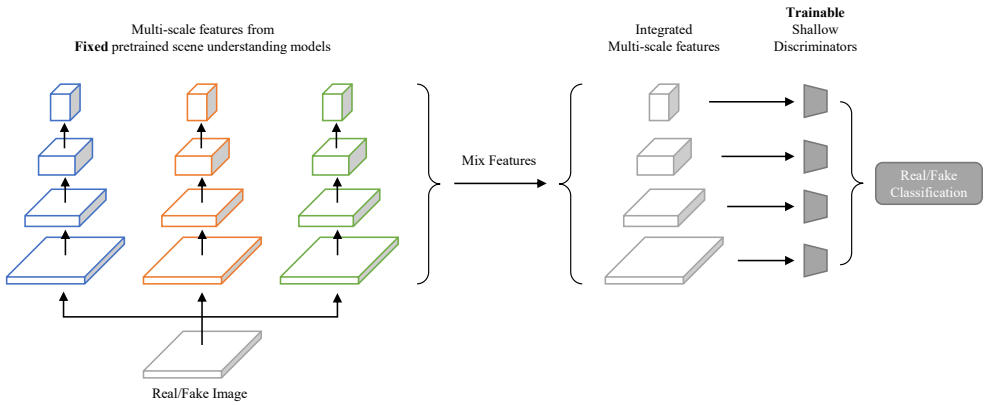


Figure 5.1 Overview of the proposed method

5.2 Method

5.2.1 Leveraging Pretrained Vision Models

To enhance the discrimination process, we employ pretrained vision models as powerful feature extractors. We first extract multi-scale features $\{f_i\}_{i=1}^{L}$ from a fixed pretrained vision model F and train shallow discriminators $\{C_i\}_{i=1}^{L}$ based on the extracted features. In addition to the original adversarial loss $V(G, D)$, the adversarial loss calculated using the pretrained models is jointly optimized to further improve the model:

$$\min_{G} \max_{D, D_{aux}} V(G, D) + V(G, D_{aux}), \tag{5.1}$$

where $D_{aux}(x) = \sum_{i=1}^{L} C_i(f_i)$ and L is the number of different scales. In this chapter, we aim to improve the complex scene generation that are challenging for GAN models. Therefore, we utilize all multi-scale features which are useful for recognition and discrimination on local regions, and accordingly utilize multiple shallow discriminators to process features of each scale.

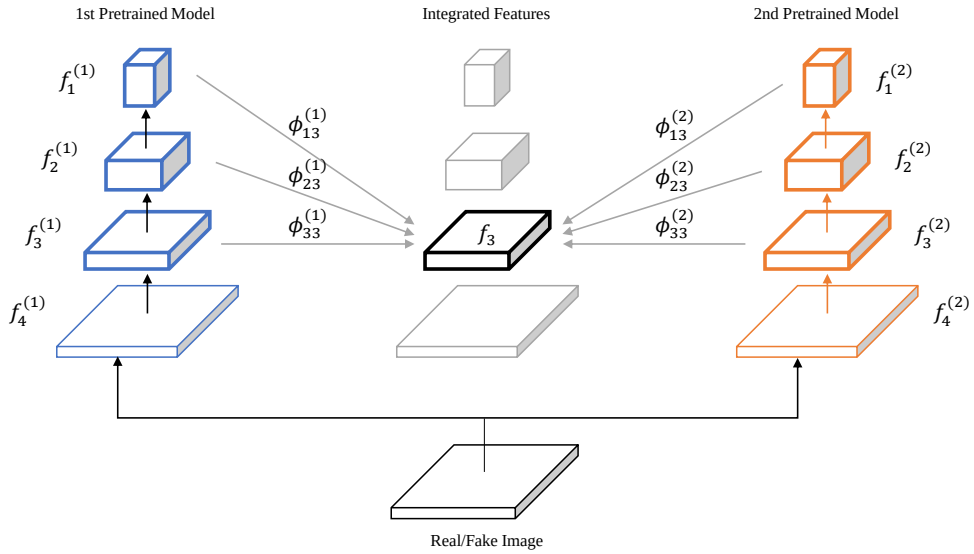


Figure 5.2 Illustration of proposed feature ensemble method

5.2.2 Feature Ensemble across Scales and across Models

Models trained on different vision tasks contain different knowledge on the images. Especially for complex scene images, there exist various scene understanding tasks that require quite different knowledge to solve the task such as object detection, semantic segmentation, and depth estimation. We aim to fully leverage these task-specific knowledge or representations to enhance discrimination process. To this end, we propose to ensemble multi-scale features extracted from multiple models across scales and across models.

Concretely, we denote the multi-scale features from k-th pretrained model as $\{f_i^{(k)}\}_{i=1}^{L}$. To ensemble features from different models, we use projection layers to transform the extracted features into integrated feature spaces. We denote a projection layer that maps j-th scale feature of k-th pretrained model

into integrated feature space of i-th scale by $\phi_{ji}^{(k)}$. The integrated feature of i-th scale can be calculated by summing all the projected features as follows:

$$f_i = \sum_k \sum_j \phi_{ji}^{(k)}(f_j^{(k)}). \tag{5.2}$$

We observe that the all pair connection across all scales is redundant since the backbone networks are already designed to reduce the spatial resolution of features. Therefore, we only utilize the projection networks that are in coarse-to-fine direction. Concretely, if we index the scales from coarse to fine-grained features in ascending order we can re-formulate the Equation 5.2 as follows:

$$f_i = \sum_k \sum_{j \leq i} \phi_{ji}^{(k)}(f_j^{(k)}). \tag{5.3}$$

Coarse-to-fine feedback connection is also successfully leveraged in Feature Pyramid Networks (Lin et al., 2017) and we extend this idea for feature ensemble for multi-scale features from multiple models in purpose of enhancing the discriminator. Figure 5.2 shows the detailed illustration of feature ensemble method.

5.3 Experiment

Datasets. We use Livingroom and Kitchen dataset from LSUN datasets (Yu et al., 2015) as the validation datasets. We choose these datasets because they contain a challenging image distribution that includes a variety of objects. Livingroom and Kitchen dataset contain 1.3 million and 2.2 million scene images, respectively. We resize all the images to 256×256 resolution.

Evaluation Metrics. For quantitative evaluation, we use Frechet inception distance (FID) (Heusel et al., 2017a), Kernel inception distance (KID) (Bińkowski et al., 2018), Precision and Recall (Kynkäänniemi et al., 2019) to assess the synthesis quality of generated images. Following the standard setup, all metrics are calculated using 50,000 fake images and all training images. In addition to scene level metrics, we also compare object level metrics since the visual quality of scene images are largely determined by the visual quality of individual objects in the scene. To assess the object level synthesis quality, we employ an object detector to detect objects in the generated scene images and compare the quality of detected object crops.

Comparison Methods. We compare our method with several recent GAN models. StyleGAN2 (Karras et al., 2020b) is one of the most successful GAN models that shows impressive synthesis quality of single object such as human faces, cars, and horses. Our model uses the same generator network and hyperparameters as StyleGAN2. UnetGAN (Schonfeld et al., 2020) is another recent GAN model that employs multi-scale unet architecture for the discriminator to improve discriminative ability of the discriminator. For UnetGAN, which is originally built on top of BigGAN (Brock et al., 2019) network, we modify the model to suit the better backbone of StyleGAN2 for fair comparison. Specifically, we change the building blocks to those of StyleGAN2 and apply global and local R1 regularization, resulting in better results than its official release. ProjectedGAN (Sauer et al., 2021) is a recent work that utilizes the ImageNet pretrained network for the discriminator to improve the synthesis quality and the convergence speed.

	Livingroom				Kitchen			
Method	FID↓	KID↓	Precision↑	Recall↑	FID↓	KID↓	Precision↑	Recall↑
UnetGAN (Schonfeld et al., 2020)	6.73	3.92	0.518	0.265	6.71	4.13	0.528	0.290
StyleGAN2 (Karras et al., 2020b)	4.64	2.22	0.512	0.268	5.10	2.58	0.530	0.305
StyleGAN2-ADA (Karras et al., 2020a)	4.95	2.34	0.507	0.267	6.47	3.62	0.484	0.272
ProjectedGAN (Sauer et al., 2021)	5.51	2.36	0.571	0.273	4.38	2.11	0.587	0.250
Ours	**1.50**	**0.54**	**0.578**	**0.461**	**1.65**	**0.70**	**0.588**	**0.478**

Table 5.1 Comparison result on scene-level metrics

Livingroom	couch		chair		potted plant		tv		vase	
	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓
UnetGAN (Schonfeld et al., 2020)	13.39	10.76	17.05	10.82	19.42	11.54	15.20	11.63	45.36	8.28
StyleGAN2 (Karras et al., 2020b)	11.21	8.06	14.58	8.64	16.09	8.14	12.22	9.02	40.19	6.19
ProjectedGAN (Sauer et al., 2021)	8.60	4.68	21.77	10.18	22.16	12.03	12.76	7.12	42.87	6.44
Ours	**2.51**	**0.96**	**6.34**	**2.73**	**5.81**	**1.68**	**5.68**	**1.42**	**10.66**	**2.14**
Kitchen	oven		chair		microwave		potted plant		refrigerator	
	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓
UnetGAN (Schonfeld et al., 2020)	12.70	6.22	23.79	14.01	17.84	10.26	26.90	12.63	23.45	14.31
StyleGAN2 (Karras et al., 2020b)	8.91	4.98	19.89	11.32	11.82	7.57	21.77	10.89	19.11	10.05
ProjectedGAN (Sauer et al., 2021)	13.78	6.01	21.74	9.42	20.06	12.29	25.95	8.24	21.55	13.60
Ours	**4.28**	**1.29**	**8.92**	**3.68**	**7.00**	**2.88**	**8.92**	**3.07**	**8.30**	**1.87**

Table 5.2 Object-level metrics for each object category

Implementation Details. We use StyleGAN2 generator and discriminator as the base GAN networks which is used to calculate basic adversarial loss $V(G, D)$. For auxiliary loss $V(G, D_{aux})$, we utilize pretrained backbone networks which are Swin-Transformer networks (Liu et al., 2021b). Since Swin-Transformer has shown impressive performance on various dense prediction tasks, most of the state-of-the-art scene understanding tasks utilize Swin-Transformer as their backbone network. For simplicity, we employ the pretrained network having the same network architecture.

For projection layers, we use a sequence of 1×1 convolution layer and upsampling layer. It is worth noting that the projection layers are not updated while training GAN as the aim of projection layers is to permute and reshape the extracted features rather than adding or modifying the feature space. We use four scales by default, i.e., $L = 4$ and the spatial resolutions of multi-scale features are $\{8, 16, 32, 64\}$ for the image resolution of 256.

5.3.1 Comparison Result

Scene-level metrics. In Table 5.1, we present the quantitative comparison result with state-of-the-art GAN models using scene level metrics. As shown in the result, our method substantially outperforms the baselines for all metrics. In terms of FID, our method achieves 67% and 52% improvements in each dataset compared to the best baseline methods. Especially, the recall is significantly improved, which demonstrates that our method can generate much diverse scene layouts than the baseline models.

Object-level metrics. Table 5.2 shows the comparison result of the object level synthesis quality. We report FID and KID of top 5 most frequent object categories in each data domain. As can be seen in the result, our method ob-

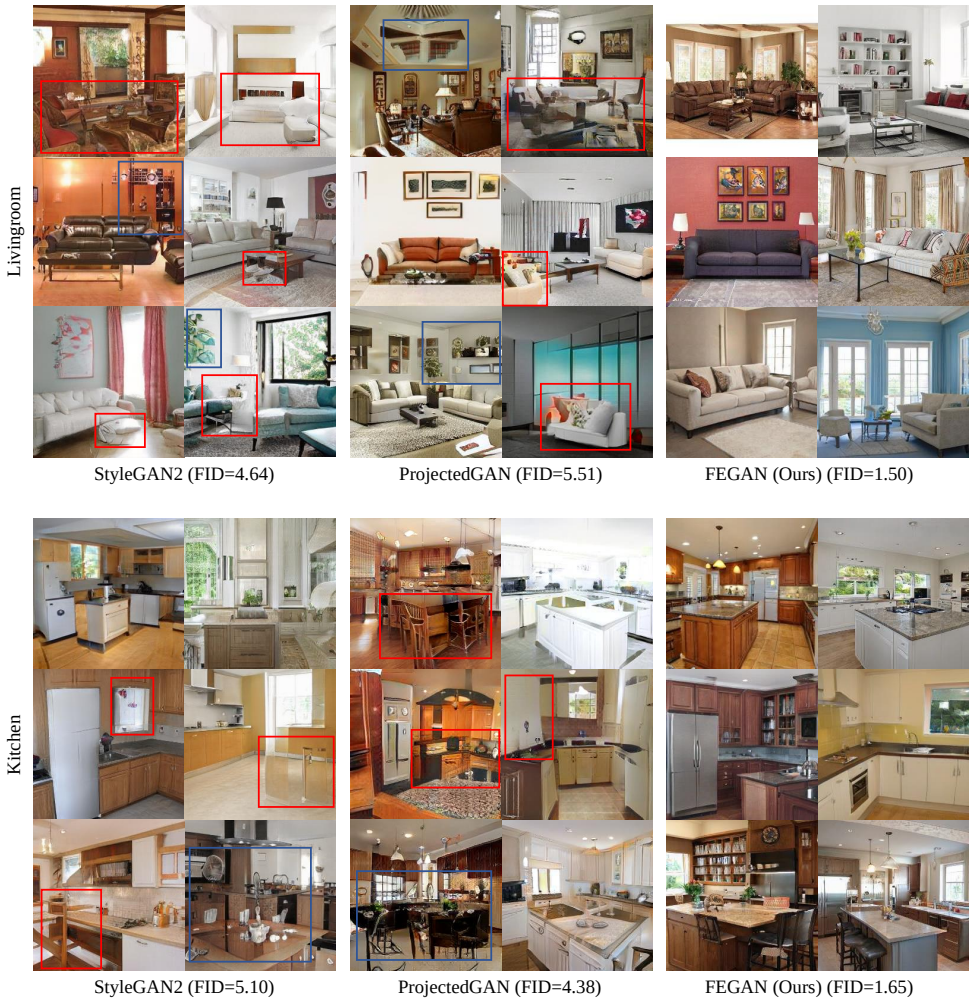


Figure 5.3 Comparison of generated samples. Red rectangles show imperfect object structures. Blue rectangles show repetitive stains making messy layouts.

tains consistently improved results for all object categories. This shows that our method effectively leverages the pretrained knowledge on various types of objects to improve the photo-realism of the objects in the generated scene images.

Qualitative comparison. Figure 5.3 shows the comparison result of samples generated by different models. All models are able to synthesize coarse layouts of indoor scenes, but the samples generated by the baselines exhibit unrealistic objects as well as mottled visual patterns. For example, objects marked with red rectangles have imperfect structure or blends unnaturally with the background. Areas inside blue rectangles contain repetitive stains that create messy layouts. By contrast, our method can generate scene images showing much clear layouts with realistic details.

5.3.2 Ablation Study

Different pretrained networks. In Table 5.3, we present the result when using different pretrained networks on Livingroom dataset. In the perspective of network architecture, transformer based networks outperform convolutional networks, e.g., EfficientNet (Tan and Le, 2019). As our target training images are complex scene images, the networks trained on object-centric images, i.e., ImageNet, are less effective than the networks trained on various scene understanding tasks. Therefore, we use pretrained networks trained on three representative scene understanding tasks of semantic segmentation (Liu et al., 2021b), object detection (Liu et al., 2021b), and depth estimation (Li et al., 2022) as feature extractors for the discriminator.

Pretrained Networks	Pretraining Dataset	FID↓	KID↓
EfficientNet (Tan and Le, 2019)	ImageNet (Deng et al., 2009)	5.51	2.36
VIT (Dosovitskiy et al., 2020)	ImageNet (Deng et al., 2009)	3.42	1.62
DINO (Dosovitskiy et al., 2020)	ImageNet (Deng et al., 2009)	3.21	1.44
Swin-T (MoBy) (Xie et al., 2021b)	ImageNet (Deng et al., 2009)	3.10	1.35
Swin-T (Segmentation) (Liu et al., 2021b)	ADE20k (Zhou et al., 2017)	2.35	1.10
Swin-T (Detection) (Liu et al., 2021b)	MS-COCO (Lin et al., 2014)	2.29	0.91
Swin-T (Depth) (Li et al., 2022)	NYU-Depth V2 (Silberman et al., 2012)	2.24	0.93
Swin-T (Seg+Det+Depth) (Ours)	-	**1.50**	**0.54**

Table 5.3 Ablation result using different pretrained networks

	Pretrained Models	W/O FE		Ours	
		FID↓	KID↓	FID↓	KID↓
Livingroom	StyleGAN2	4.64	2.22	4.64	2.22
	+ Segmentation	2.35	1.10	2.35	1.10
	+ Detection	2.25	1.04	1.95	0.86
	+ Depth Estimation	2.20	0.93	**1.50**	**0.54**
Kitchen	StyleGAN2	5.10	2.58	5.10	2.58
	+ Segmentation	2.61	1.58	2.61	1.58
	+ Detection	2.48	1.32	2.04	0.96
	+ Depth Estimation	2.40	1.27	**1.65**	**0.70**

Table 5.4 Effectiveness of feature ensemble

Effectiveness of feature ensemble. In Table 5.4, we compare the proposed method with and without feature level ensemble strategy. Without feature level ensemble, the extracted features from different models are processed independently via separate shallow discriminators to compute adversarial losses. We report the generation performance when pretrained networks are cumulatively added one by one. As can be seen in the table, the performance gain is substantially reduced without feature level ensemble, which demonstrates that the proposed feature level ensemble plays an crucial role in leveraging multi-scale features from multiple pretrained networks.

5.4 Chapter Summary

This paper proposes a method to utilize pretrained scene understanding models for complex scene generation. Based on empirical observation that the networks trained on scene understanding tasks can provide useful representations for scene generation, we propose to leverage pretrained features from multiple

models via feature level ensemble. Quantitative and qualitative experimental results show that the proposed method can synthesize diverse and realistic scene images, achieving meaningful improvements compared to recent GAN model on challenging scene datasets.

Chapter 6

Conclusion & Future Work

In this dissertation, we study ways to improve complex image generation by enhancing the discriminator for GAN training. To enhance the discriminative ability of discriminator, we explore the methods for improvement in three different directions: (1) designing a better architecture, (2) designing a new auxiliary task for the discriminator, and (3) designing a transfer learning method to leverage pretrained vision models.

In Chapter 3, we propose an advanced architecture for conditional image generation task designed to process multi-label condition. We introduce conditional modules for each of the generator and discriminator, named *ADGAN*. For generator part, we propose a product-of-Gaussian based conditional prior sampling method. To encode a set of attribute labels, we map the input labels into corresponding Gaussian distributions and aggregate them to a unified distribution by product-of-Gaussian calculation. For discriminator, we introduce an attention-based conditional discriminator to improve the discrimination process for given multi-label condition. To do this, we use the attention mechanism

to acknowledge the discriminator which area to focus on to better discriminate the image for a given label. We validate our method on fashion image dataset and show proposed ADGAN outperforms the baseline method, providing better controllability on image generation task.

In Chapter 4, we design an auxiliary task which is assigned to discriminator to improve its discriminative ability, therefore better model complex image distributions such as scenes with multiple objects. Considering the structural complexity of scene images containing multiple local objects, we build a multi-scale discriminator that produces loss signals not only for global representation but also local representations of multiple scales. To further enhance the local-to-global representations, we design a multi-scale contrastive learning task and assign the task to the discriminator so that the original binary classification task benefits from the enhanced representations. To impose spatial consistency on contrastive learning for local representations, we identify the positive and negative feature pairs using predefined distance threshold where the distance is computed between the pixel coordinates of local features. From experiments on challenging scene image datasets, we observed that the proposed auxiliary task substantially improves the synthesis quality of generated scene images in terms of both scene-level and object-level metrics. Through extensive ablation study, we validated the efficacy of local representation learning and distance thresholding strategy under various model configurations. While we verify the proposed auxiliary task on unconditional generation setup, we also expect our finding could be applied to other conditional image generation tasks since our method operates in a self-supervised manner.

In Chapter 5, we explore the use of pretrained vision models to aid the discriminator in learning complex structural difference between real and fake scene images. We assume that the models trained on various scene understand-

ing tasks contain useful knowledge which is helpful for discriminative task of GAN training. Therefore, we sought for improvement of generation quality using pretrained scene understanding models as powerful feature extractors. Since the models trained on different tasks contain distinct knowledge on scenes, we propose to ensemble the features extracted from multiple models to form a set of unified multi-scale features, which are then used to discriminate real and fake images. The feature-level ensemble strategy enables efficient utilization of common and distinct knowledge learned by different models thus further boosts the synthesis quality as well as diversity of generated scene images. Experiments on two challenging indoor scene dataset show the proposed method significantly lower the FID and KID score in both scene-level and object-level evaluation. As an ablation study, we compare the result when using different pretrained vision models trained for different tasks involving different datasets. The result shows that both network architecture and pretraining task affects the usefulness of pretrained representations, where the models trained on scene understanding tasks and scene image datasets learns the most useful representations. We also validate the proposed feature ensemble strategy by comparing the results with or without feature ensemble and we observed meaningful improvement is achieved by feature level ensemble.

Generative models have a great potential on various synthesis and editing tasks as well as generative data augmentation for various recognition tasks. Although recent advances of GAN models enabled photo-realistic synthesis of various objects, there still exist challenges on modeling more complex image distributions. We believe that the methods discussed in this dissertation would help alleviating the burden of discriminator and making it more robust, consequently improving the generation of complex images.

As the future work, we consider three research directions to extend our ap-

proach. First, while this dissertation consider generative models only for image data, we could apply our findings to generative models for other data types such as speech and video. For example, a recent work (Skorokhodov et al., 2022) has proposed a generative adversarial network for generating long videos based on StyleGAN architecture. We believe the performance could be further boosted by leveraging methods proposed in this dissertation.

Since we have investigated the orthogonal directions for improving discriminator learning, the proposed methods can be utilized together to further boost the generation performance. For examples, the self-supervised auxiliary task described in Chapter 4 can be integrated to the feature ensemble process in Chapter 5. While the multi-scale contrastive learning learns beneficial representations from stochastic image augmentations, the FEGAN does not utilize such information, therefore the performance can be further improved by leveraging both techniques.

Lastly, we believe the strategies proposed in this dissertation can be also applied to other types of generative models such as diffusion probabilistic models (Ho et al., 2020). While different model classes have distinct mechanism to learn underlying data distribution, the high-level strategies investigated in this dissertation can be applied commonly. For example, pretrained representations learned by various vision tasks would be also useful for learning complex data distributions in unconditional diffusion models. We leave these research themes as our future work.

Chapter 7

Appendix

7.1 Detailed Network Architecture

7.1.1 Network Architecture of ADGAN

Table 7.1 and Table 7.2 show the architectural details of generator and discriminator of ADGAN, respectively. The basic blocks of discriminator and generator are similar to network used in SNGAN (Miyato et al., 2018). We can apply attention block to one or more layers of the discriminator. Here we use only one attention block.

7.1.2 Network Architecture of MsConD

Table 7.3 shows the architectural details of MsConD discriminator. Our discriminator is built upon the backbone resnet-based discriminator used in StyleGAN2 (Karras et al., 2020b). We use branches to process the feature map at each level, where each branch consists of three components: a shared block, a discriminator head and a projection head. The shared block consists of 3 1×1

convolutional layers with residual connections. The shared block translates a feature map in the backbone network into an intermediate feature map of the same size. The intermediate feature map is then projected into two different outputs each by a discriminator head and a projection head, where both head layers are implemented with 1×1 convolutional layers. The discriminator head is a single convolutional layer, while the projection head consists of two convolutional layers, i.e., Conv-ReLU-Conv. We set the channel dimension of the projection output, i.e., C_p, as 256. We use ReLU activation for all layers in branches.

7.2 Additional Samples

7.2.1 Comparison of additional samples of MsConD

For comparisons to the state of the art models, we provide more uncurated samples generated by different models in Figure 7.1, 7.2, 7.3. Compared to baselines, MsConD produces convincing results of more realistic scene images with improved local details.

Layer	Kernel	Output Shape
SNConv, lRelu	[3,3,1]	$h \times w \times 64$
SNConv, lRelu	[4,4,2]	$\frac{h}{2} \times \frac{w}{2} \times 128$
SNConv, lRelu	[4,4,2]	$\frac{h}{4} \times \frac{w}{4} \times 256$
SNConv, lRelu	[3,3,1]	$\frac{h}{4} \times \frac{w}{4} \times 256$
Attention	-	$\frac{h}{4} \times \frac{w}{4} \times 512$
SNConv, lRelu	[4,4,2]	$\frac{h}{8} \times \frac{w}{8} \times 512$
SNConv, lRelu	[3,3,1]	$\frac{h}{8} \times \frac{w}{8} \times 512$
Linear	-	1

Table 7.1 Discriminator Architecture of ADGAN

Layer	Kernel	Output Shape
z	-	256
Linear, BN, ReLU	[3,3,1]	$\frac{h}{16} \times \frac{w}{16} \times 512$
Upsample	-	$\frac{h}{8} \times \frac{w}{8} \times 512$
Conv, BN, ReLU	[3,3,1]	$\frac{h}{8} \times \frac{w}{8} \times 256$
Upsample	-	$\frac{h}{4} \times \frac{w}{4} \times 256$
Conv, BN, ReLU	[3,3,1]	$\frac{h}{4} \times \frac{w}{4} \times 128$
Upsample	-	$\frac{h}{2} \times \frac{w}{2} \times 128$
Conv, BN, ReLU	[3,3,1]	$\frac{h}{2} \times \frac{w}{2} \times 64$
Upsample	-	$h \times w \times 64$
Conv, BN, ReLU	[3,3,1]	$h \times w \times 32$
Conv, Tanh	[3,3,1]	$h \times w \times 3$

Table 7.2 Generator Architecture of ADGAN

Type	Layer	Output Shape
-	Image	$256 \times 256 \times 3$
backbone	ResBlk	$256 \times 256 \times 128$
backbone	ResBlk	$128 \times 128 \times 256$
backbone	ResBlk	$64 \times 64 \times 512$
backbone	ResBlk	$32 \times 32 \times 512$
backbone	ResBlk	$16 \times 16 \times 512$
branch: shared	1×1 ResBlk	$16 \times 16 \times 512$
branch: disc	1×1 Conv	$16 \times 16 \times 1$
branch: proj	1×1 Conv	$16 \times 16 \times 256$
backbone	ResBlk	$8 \times 8 \times 512$
branch: shared	1×1 ResBlk	$8 \times 8 \times 512$
branch: disc	1×1 Conv	$8 \times 8 \times 1$
branch: proj	1×1 Conv	$8 \times 8 \times 256$
backbone	ResBlk	$4 \times 4 \times 512$
backbone	Flatten	8192
backbone	Linear	512
branch: disc	Linear	1
branch: proj	Linear	256

Table 7.3 Discriminator Architecture of MsConD

StyleGAN2 (FID 8.04)

ProjectedGAN (FID 5.07)

InsGen (FID 4.21)

MsConD (Ours) (FID 2.63)

Figure 7.1 Uncurated Samples for Cityscapes

StyleGAN2 (FID 4.64)

ProjectedGAN (FID 5.51)

InsGen (FID 4.17)

MsConD (Ours) (FID 2.73)

Figure 7.2 Uncurated Samples for Livingroom

StyleGAN2 (FID 5.10)

ProjectedGAN (FID 4.38)

InsGen (FID 5.76)

MsConD (Ours) (FID 2.88)

Figure 7.3 Uncurated Samples for Kitchen

Bibliography

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Arantxa Casanova, Michal Drozdzal, and Adriana Romero-Soriano. Generating unseen complex scenes: are we there yet? *arXiv preprint arXiv:2012.04027*, 2020.

Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12154–12163, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.

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

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

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

Raghudeep Gadde, Qianli Feng, and Aleix M Martinez. Detail me more: Improving gan's photo-realism of complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13950–13959, 2021.

Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2, 2014.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016b.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum

contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

Xiaodong He and Li Deng. Deep learning for image-to-text generation: A technical overview. *IEEE Signal Processing Magazine*, 34(6):109–116, 2017.

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. *Advances in neural information processing systems*, 30, 2017a.

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. *Advances in neural information processing systems*, 30, 2017b.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Liang Hou, Huawei Shen, Qi Cao, and Xueqi Cheng. Self-supervised GANs with label augmentation. In *Advances in Neural Information Processing Systems*, 2021.

Tianyu Hua, Hongdong Zheng, Yalong Bai, Wei Zhang, Xiao-Ping Zhang, and Tao Mei. Exploiting relationship for complex-scene image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1584–1592, 2021.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

Jongheon Jeong and Jinwoo Shin. Training gans with stronger augmentations via contrastive discriminator. In *International Conference on Learning Representations*, 2021.

Shuhui Jiang and Yun Fu. Fashion style generator. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3721–3727. AAAI Press, 2017.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020b.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In *International Conference on Machine Learning (ICML)*, pages 595–603, 2014.

Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10651–10662, 2022.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32:3927–3936, 2019.

Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021.

Fahad Lateef and Yassine Ruichek. Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 338:321–348, 2019.

Hanbit Lee and Sang-goo Lee. Fashion attributes-to-image synthesis using attention-based generative adversarial network. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 462–470. IEEE, 2019.

Hanbit Lee, Youna Kim, and Sang-goo Lee. Multi-scale contrastive learning for complex scene generation. In *2023 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2023.

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022.

Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017a.

Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized {gan} training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*, 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021b.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. *Advances in neural information processing systems*, 30, 2017.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018a.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018b.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014a.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014b.

Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *International Conference on Learning Representations*, 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Yao Ni, Piotr Koniusz, Richard Hartley, and Richard Nock. Manifold learning benefits gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11265–11274, 2022.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 2021.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

Pedro O Pinheiro, Amjad Almahairi, Ryan Y Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. In *NeurIPS*, 2020.

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

Scott Reed, Aäron van den Oord, Nal Kalchbrenner, Victor Bapst, Matt Botvinick, and Nando de Freitas. Generating interpretable images with controllable structure. *Technical report*, 2016.

Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1153, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. *Advances in neural information processing systems*, 30, 2017.

Ruslan Salakhutdinov and Hugo Larochelle. Efficient learning of deep boltzmann machines. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 693–700. JMLR Workshop and Conference Proceedings, 2010.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2234–2242, 2016.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Othman Sbai, Mohamed Elhoseiny, Antoine Bordes, Yann LeCun, and Camille Couprie. Design: Design inspiration from generative networks. *arXiv preprint arXiv:1804.00921*, 2018.

Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8207–8216, 2020.

Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *International Conference on Learning Representations*, 2021.

Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole,

and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International workshop on simulation and synthesis in medical imaging*, pages 1–11. Springer, 2018.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3626–3636, 2022.

Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2647–2655, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16515–16525, 2022.

Ngoc-Trung Tran, Viet-Hung Tran, Bao-Ngoc Nguyen, Linxiao Yang, and Ngai-Man Man Cheung. Self-supervised gan: Analysis and improvement with multi-class minimax game. *Advances in Neural Information Processing Systems*, 32:13253–13264, 2019.

Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7921–7931, 2021.

Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.

Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. In *International Conference on Learning Representations (ICLR)*, 2018.

Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks. *arXiv preprint arXiv:2105.04206*, 2021a.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense

contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021b.

Ryan Webster, Julien Rabin, Loic Simon, and Frédéric Jurie. Detecting overfitting of deep generative networks via latent recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11273–11282, 2019.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *International Conference on Machine Learning (ICML)*, 2018.

Yue Wu, Pan Zhou, Andrew G Wilson, Eric Xing, and Zhiting Hu. Improving gan training with probability ratio clipping and sample reweighting. *Advances in Neural Information Processing Systems*, 33:5729–5740, 2020.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. *arXiv preprint arXiv:2103.12902*, 2021.

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021a.

Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021b.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021c.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *arXiv preprint*, 2017.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Data-efficient instance generation from instance discrimination. *Advances in Neural Information Processing Systems*, 34:9378–9390, 2021.

Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tiangchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. In *International Conference on Learning Representations*, 2019.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked

attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.

Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2327–2336, 2019.

Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. Data augmentation for spoken language understanding via joint variational generation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7402–7409, 2019.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019.

Ning Yu, Guilin Liu, Aysegul Dundar, Andrew Tao, Bryan Catanzaro, Larry S Davis, and Mario Fritz. Dual contrastive loss and attention for gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6731–6742, 2021.

Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault. The ai index 2022 annual report. `https://aiindex.stanford.edu/wp-content/`

uploads/2022/03/2022-AI-Index-Report_Master.pdf, 2022. Accessed: 2022-10-30.

Han Zhang, Tao Xu, and Hongsheng Li. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5908–5916. IEEE, 2017.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations*, 2020.

Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020a.

Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. *arXiv preprint arXiv:2006.02595*, 2020b.

Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for gans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11033–11041, 2021.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017a.

Shizhan Zhu, Sanja Fidler, Raquel Urtasun, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. *arXiv preprint arXiv:1710.07346*, 2017b.

Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019.

초록

Generative Adversarial Network (GAN) 은 최근 몇 년 동안 가장 성공적인 생성 모델 중 하나이다. 생성기와 판별기를 사용한 적대적 훈련 방식은 이미지와 같은 고차원 데이터 분포를 모델링하는 새롭고 강력한 방법을 제공한다. 이 메커니즘에서 판별기의 판별 기능은 핵심적인 역할을 한다. 이는 생성기가 판별기가 실제 샘플과 가짜 샘플을 구별해낼 수 있는 능력에 전적으로 의존하여 생성기의 생성 성능을 향상 시킬 수 있기 때문이다. 본 논문에서는 이러한 판별기 학습을 향상시키기 위한 세 가지 방법을 제안함으로써 GAN 모델의 개선을 모색하였다.

먼저, 복잡한 다중 레이블 조건에 대한 조건부 이미지 생성을 개선하기 위해 Attention-based Discriminator (ADGAN)을 제안한다. ADGAN은 판별기가 조건 레이블과 관련된 이미지 영역에 집중할 수 있도록 Attention 기법을 활용하는 판별기를 제안한다. 또한 다중 레이블 조건을 효율적으로 인코딩하기 위해 가우시안 곱 기반의 잠재 벡터 샘플링 방법을 제안한다. 제안된 아키텍처는 복잡하고 다양한 속성 라벨에 대해 이미지 생성 프로세스의 제어 가능성을 향상시켰다.

그 다음으로 여러 객체가 있는 장면 이미지와 같이 보다 복잡한 이미지에 대한 판별기 향상 방법에 대해 성능 향상을 모색한다. 장면 이미지는 일반적으로 이미지의 구조적 복잡성이 높기 때문에 판별기가 실제 장면 이미지와 가짜 장면 이미지 간의 복잡한 구조적 차이를 구별해야 하기에 학습의 난이도가 높다. 우리는 판별기의 학습을 돕기 위해 판별기의 로컬 표현을 향상키기기 위한 다중 스케일 대조 학습 (Multi-scale Contrastive Learning)을 설계하고 이를 통해 판별기에 추가 작업으로 부여한다. 이를 통해 이미지의 로컬 구조에 대한 판별기의 판별 능력을 강화하여 결과적으로 장면 생성 성능을 향상시킬 수 있었다.

마지막으로 장면 생성 성능을 더 향상시키기 위해 사전 훈련된 장면 이해 모델을 활용하여 판별기의 판별 과정을 추가로 지원하는 방법을 탐색합니다. 사전

훈련된 모델들은 장면 이미지의 복잡한 의미 구조에 대한 풍부한 지식을 학습하고 있으므로 사전 훈련된 표현을 사용하여 판별기의 판별 능력을 증진시켰다. 여러 전문 모델이 담고 있는 공통적인 지식과 모델 별 고유한 지식을 최대한 활용하기 위해 모델로부터 추출된 피쳐들을 앙상블하여 통합된 다중 스케일 기능 세트를 형성하고 이를 판별 과정에 활용할 것을 제안한다.

우리는 다양한 이미지 도메인에서의 성능 평가 및 분석을 통해 제안된 방법들이 복잡한 이미지 분포 모델링에서 의미 있는 개선을 달성하였다. 이러한 성과가 생성 모델의 문제와 한계를 극복하고 생성모델의 다운스트림 애플리케이션들도 용이하게 하는 데 도움이 될 것으로 기대한다.

주요어: 생성적 적대 신경망, 심층 생성 모델, 이미지 생성, 조건부 이미지 생성, 판별기 강화, 장면 생성, 자기 지도 학습, 전이 학습

학번: 2013-20865