High-Quality and Diverse Few-Shot Image Generation via Masked Discrimination

Jingyuan Zhu, Huimin Ma, Senior Member, IEEE, Jiansheng Chen, Senior Member, IEEE, and Jian Yuan

Abstract—Few-shot image generation aims to generate images of high quality and great diversity with limited data. However, it is difficult for modern GANs to avoid overfitting when trained on only a few images. The discriminator can easily remember all the training samples and guide the generator to replicate them, leading to severe diversity degradation. Several methods have been proposed to relieve overfitting by adapting GANs pre-trained on large source domains to target domains using limited real samples. This work presents masked discrimination to realize few-shot GAN adaptation, which is the first feature-level augmentation method for generative tasks. Random masks are applied to features extracted by the discriminator from input images. We aim to encourage the discriminator to judge various images that share partially common features with training samples as realistic. Correspondingly, the generator is guided to generate diverse images instead of replicating training samples. In addition, we employ a cross-domain consistency loss for the discriminator to keep relative distances between generated samples in its feature space. It strengthens global image discrimination and guides adapted GANs to preserve more information learned from source domains for higher image quality, resulting in better cross-domain correspondence. The effectiveness of our approach is demonstrated both qualitatively and quantitatively with higher quality and greater diversity on a series of few-shot image generation tasks than prior methods.

Index Terms—Few-shot image generation, masked discrimination, cross-domain correspondence.

I. INTRODUCTION

Modern generative adversarial networks (GANs) [1] containing abundant parameters have achieved great success in image generation with large amounts of training data. For example, BigGAN [2] uses more than 100M parameters to achieve significant improvement on ImageNet. StyleGAN2 [3] is trained on 70,000 images from FFHQ [4] with a total of 26.2M trainable parameters in its generator. However, GANs easily overfit and tend to replicate training data instead of generating diverse samples when trained on datasets containing fewer samples [5]. Unfortunately, only a few samples can be obtained in some corner cases like artists’ paintings.

For few-shot image generation tasks, the complex discriminator can remember all the training samples and excessively guide the generator to replicate them, leading to overfitting and serious diversity degradation. Several prior works [6], [7], [8], [9], [10], [11] have been proposed to adapt GANs pre-trained on related large source domains to target domains with limited data. They aim to preserve the diversity of adapted GANs by keeping information learned from source domains. Data augmentation approaches [12], [13] also play positive roles in few-shot image generation. However, these approaches mostly need hundreds of training samples from target domains to produce high-quality results. When trained on extremely few samples (e.g., 10 images), they cannot avoid overfitting or replicating training samples, resulting in limited quality and diversity.

Recent works [14], [15] propose to build a one-to-one correspondence between the source and target domains, leading to greater diversity. A cross-domain consistency loss designed for the generator is introduced in Cross-domain Consistency (CDC) [14] to preserve pairwise distance information learned from source domains during adaptation. Moreover, CDC adds patch-level discrimination to relieve overfitting. The patch-level discrimination judges realism based on intermediate features of the discriminator, where the receptive field of each member corresponds to a patch in the input image. Nevertheless, CDC still fails to preserve the diversity of some detailed characteristics like detailed hairstyles and facial expressions of humans, resulting in limited diversity.

This work focuses on regulating the discriminator, which easily overfits and guides the generator to replicate training samples in few-shot image generation tasks. As shown in Fig. 1, our approach guides adapted GANs to learn from target domains with limited data and allows them to preserve the information learned from source domains, improving generation quality and diversity. We propose masked discrimination to further relieve overfitting with feature-level augmentation and achieve more diverse results. We apply random masks to features extracted by the target discriminator, forcing it to judge realism with partially masked features. As a result, the target generator is guided to generate various images sharing partially common features with real samples, leading to greater diversity. Besides, we adapt the cross-domain consistency loss [14] to discriminators, aiming to achieve higher quality by strengthening global image discrimination and preserving information learned from source domains. It encourages the
target discriminator to extract diverse features from input images and keep relative distances between generated samples in its feature space during domain adaptation.

The main contributions of this paper are concluded as: (1) We propose masked discrimination, which is the first feature-level augmentation method for generative tasks and guides adapted GANs to generate images sharing partially common features with training samples and achieve greater diversity. (2) We present the discriminator cross-domain consistency loss to preserve information learned from source domains and achieve higher quality. (3) We demonstrate the effectiveness of our approach on a series of few-shot image generation tasks with qualitative and quantitative results outperforming prior works.

II. RELATED WORK

Few-shot image generation Few-shot image generation aims to generate diverse and high-quality images using only a few available samples. Most existing works follow the adaptation method proposed in TGAN [9] to adapt GANs pre-trained on large source domains, including ImageNet [16], FFHQ [4], and LSUN [17] et al., to target domains with limited data. Augmentation approaches [13], [18], [19], [20] like ADA [12] can provide more augmented samples to relieve overfitting. BSA [8] updates the scale and shift parameters in the generator only during adaptation. FreezeD [7] freezes the high-resolution layers of the discriminator to relieve overfitting. EWC [6] applies elastic weight consolidation to regularize the generator by making it harder to change critical weights that have higher Fisher information [21] values during adaptation. MineGAN [10] adds additional fully connected networks to modify noise inputs for the generator, aiming to shift the distributions of latent space for adaptation. CDC [14] builds a correspondence between source and target domains with the cross-domain consistency loss for generators and patch-level discrimination. DCL [15] maximizes the similarity between images in the source and target domains and pushes away the generated samples from real images for greater diversity. LFS-GAN [22] focuses on lifelong few-shot image generation. DDPM-PA [23] first investigates few-shot image generation with diffusion models. Our approach is compared with the abovementioned approaches in quality and diversity to prove its effectiveness on few-shot training data.

In addition, recent works in few-shot image generation have provided different research perspectives. RSSA [24] proposes a relaxed spatial structural alignment method with compressed latent space based on inverted GANs [25]. RSSA aims to preserve image structures learned from source domains, which is inappropriate for some abstract target domains like artists’ paintings. AdAM [26] and RICK [27] introduce domain adaptation approaches that work especially well for unrelated source/target domains. Recent works including OSG [28], MTG [29], OSCLIP [30], GDA [31], and DIFA [32], et al. focus on exploring single-shot GAN adaptation with the help of pre-trained CLIP [33] image encoders.

Domain translation Domain translation research has proposed many approaches based on conditional GANs [34], [35], [36] and variational autoencoders (VAEs) [37] to convert an image from source domains to target domains. However, most domain translation approaches [38], [39] ask for much training data in both source and target domains, constraining its application to few-shot tasks. Recent works [40], [41] separate the content and style in image translation to relieve this problem. SEMIT [42] applies semi-supervised learning with a noise-tolerant pseudo-labeling procedure. However, abundant class or style-labeled data are still required for training. This paper aims to realize model-level unconditional few-shot image generation based on models pre-trained on large-scale source domains and only a few real samples instead of image-level translation.

III. APPROACH

Few-shot image generation aims to achieve high quality and great diversity generation with a few real samples utilizing the source GAN model, which consists of the source generator $G_s$ and source discriminator $D_s$. However, adapted GANs directly fine-tuned on limited data overfit seriously and tend to replicate real samples since the target discriminator $D_t$ can remember all of them with its deep neural networks. Given $x$ representing real samples following data distributions $p_{data}(x)$ and $z$ representing random noise inputs, the non-saturating loss for directly fine-tuned GANs is given by:

$$L_{adv}(G_t, D_t) = D_t(G_t(z)) - D_t(x).$$

Different from existing methods designed for generators [6], [10], [22], [24], our approach focuses on adjusting the optimization target of adapted GANs by regulating the target discriminator $D_t$. We follow CDC [14] to add patch-level discrimination, which computes adversarial loss using certain intermediate features of the target discriminator $D_t$. The full target discriminator $D_t$ is used for noises sampled from a subset of the latent space $Z_{sub}$ and the patch-level discrimination is applied to the whole latent space.

The main contribution of our approach is to propose masked discrimination, which further relieves overfitting for discriminators. More specifically, we propose to apply random masks to features extracted by the target discriminator $D_t$. The target discriminator $D_t$ is encouraged to judge various images sharing partially common features with training samples as realistic for greater generation diversity (Sec III-A). To further preserve information learned from source domains, we adapt the cross-domain consistency loss (CDC loss) [14], which is originally designed for generators, to regularize the target discriminator $D_t$ for higher generation quality (Sec III-B). The target generator $G_t$ is trained to learn from the limited real samples and preserve diversity provided by pre-trained source GANs under the guidance of the target discriminator $D_t$.

A. Masked Discrimination

Motivation The target discriminator $D_t$ easily overfits since it can remember all the training samples and guide the target generator $G_t$ to replicate them in few-shot image generation tasks. However, adapted models should extract the common features of target domains from limited data and
Based on a GAN model ($G_s + D_s$) pre-trained on a large source domain (e.g., FFHQ), we propose to adapt it to target domains using limited data (e.g., 10 images) via masked discrimination (adapted GANs: $G_t + D_t$). Our approach adapts the pre-trained source model to target domains naturally and maintains generation diversity with better cross-domain correspondence.

We propose a simple but effective approach to apply random masks to features extracted by the target discriminator $D_t$ from input fake generated images $G_t(z)$ or real images $x$ as shown in Fig. 2. In this way, $D_t$ is guided to focus on the common features that have greater possibilities to be preserved after random masking. Besides, it becomes difficult for $D_t$ to remember any training samples since it has no access to complete features extracted from few-shot data. Fake images containing the common features of target domains can be judged as real, even if they are very different from training samples. Therefore, the target generator $G_t$ can produce more diverse images of target domains, slowing down diversity degradation during few-shot domain adaptation.

**Method** We use the StyleGAN2 [3] network architecture as an example and achieve the most realistic results with great diversity when masking the output features of discriminator block 4, as shown in Fig. 2. Besides, we use different sizes of masks ranging from 1/2 to 3/4 of the feature space size and produce pleasing results as illustrated in Sec IV. The masks are not patch-level but completely random. We visualize output features as a group of cubes in Fig. 2 to show the mask ratio clearly. Detailed ablations of masks are provided in Sec. IV-C.

**Novelty** Masked discrimination applies random masks to the output features of an intermediate layer in the discriminator. It is the first feature-level augmentation method applied to generative tasks. It differs from existing augmentation approaches [12], [13], [18], [19], [20] applied at the image level, which aims to provide some additional training samples to relieve overfitting and patch-shuffle regularization [43] which creates rich local variations for images to relieve overfitting. Our approach is designed to extract the common features of target domains from extremely limited training data instead of replicating them. It is also different from the dropout [44], [45] method, which randomly stops updating part of neurons during training. Masked discrimination is a novel and effective approach to relieve overfitting for GANs trained on limited data.

**B. Discriminator CDC Loss**

**Motivation** While masked discrimination improves the generation diversity of adapted GANs, it weakens global image discrimination and leads to degraded generation quality since the target discriminator $D_t$ has no access to complete features extracted from input images. Therefore, we propose discriminator CDC loss to preserve more information learned from source domains for higher quality and better visual effects.
Method As shown in Fig. 3, we propose to apply the cross-domain consistency loss [14] to both the generator and discriminator, aiming to keep the pairwise relative distances between generated samples and preserve more information provided by source GANs for higher image quality. Given two arbitrary noise inputs $z_i, z_j$ for the generator, the pairwise relative distances in the feature space of the source and target generator can be given by $\text{sim}(G^m_s(z_i), G^m_t(z_j))$ and $\text{sim}(G^m_t(z_i), G^m_t(z_j))$, where $\text{sim}$ represents the cosine similarity between activations at the $m$th layer of the generators. Similarly, we have pairwise relative distances in the feature space of the source and target discriminator as $\text{sim}(D^p_s(G_s(z_i)), D^p_t(G_t(z_j)))$ and $\text{sim}(D^p_t(G_t(z_i)), D^p_t(G_t(z_j)))$, where $\text{sim}$ represents the cosine similarity between activations at the $n$th layer of the discriminators. A batch of $K + 1$ noise vectors $\{z_k\}_{k=0}^{K}$, $(0 \leq k \leq K)$ is needed to construct $K$-way probability distributions for an arbitrary noise vector $z_i$ in the source and target generator as follows:

$$p^{s,m}_{g,i} = \text{Softmax}(\{\text{sim}(G^m_s(z_i), G^m_t(z))\}_{j \neq i})$$

$$p^{t,m}_{g,i} = \text{Softmax}(\{\text{sim}(G^m_t(z_i), G^m_t(z))\}_{j \neq i}).$$  \hspace{1cm} (2)

Similarly, the probability distributions in the source and target discriminator are as follows:

$$p^{s,n}_{d,i} = \text{Softmax}(\{\text{sim}(D^n_s(G_s(z_i)), D^n_t(G_t(z))\}_{j \neq i})$$

$$p^{t,n}_{d,i} = \text{Softmax}(\{\text{sim}(D^n_t(G_t(z_i)), D^n_t(G_t(z))\}_{j \neq i})$$ \hspace{1cm} (3)

Finally, we have the cross-domain consistency loss $L_{dist}$ across layers and image instances using KL-divergence ($D_{KL}$) for the generator and discriminator as:

$$L_{dist}(G_s, G_t) = \mathbb{E}_{z \sim p_{z}(z)} \sum_{m,i} D_{KL}(p^{t,m}_{g,i} || p^{s,m}_{g,i}),$$

$$L_{dist}(D_s, D_t) = \mathbb{E}_{z \sim p_{z}(z)} \sum_{n,i} D_{KL}(p^{t,n}_{d,i} || p^{s,n}_{d,i}).$$  \hspace{1cm} (4)

where $p_{z}(z)$ represents the distributions of noise inputs. We aim to encourage the target discriminator to extract diverse features like the source discriminator instead of only focusing on features of a few real samples. Correspondingly, the generator is guided to preserve more detailed characteristics learned from source domains during adaptation. Our full approach combines discriminator CDC loss with masked discrimination to achieve more diverse adapted samples and preserve more information from source GANs naturally.

**Novelty** CDC loss [14] is designed to regularize the target generator $G_t$ by building correspondence between source and target domains. We find that building CDC loss based on the features in source and target discriminators is an effective approach matching with the masked discrimination approach to reduce blurs and artifacts in adapted samples and improve generation quality. It strengthens global image discrimination by preserving more information from source samples.

C. Overall Optimization Target

The overall loss function of the proposed approach consists of the adversarial loss using image-level and patch-level discrimination and cross-domain consistency loss for the target generator and discriminator:

$$L = \mathbb{E}_{x \sim p_{data}(x)}[\mathbb{E}_{z \sim p_{z}(z)} L_{adv}(G_t, D_t)$$

$$+ \mathbb{E}_{z \sim p_{z}(z)} L_{adv}(G_t, D_p)]$$

$$+ \lambda (L_{dist}(G_s, G_t) + L_{dist}(D_s, D_t)).$$  \hspace{1cm} (5)

Here $D_p$ represents a subset of the target discriminator $D_t$ used for patch-level discrimination. $D_p$ uses intermediate features of $D_t$, where the receptive field of each member corresponds to a patch in the input image. The proposed masked discrimination has influences on both image-level and patch-level discrimination. We find $\lambda$ between $10^3$ to $5 \times 10^3$ appropriate for adaptation setups in this paper empirically.

IV. EXPERIMENTS

**Basic setups** We follow the experimental setups used in prior works [6], [7], [10], [14], [15] to implement the proposed approach based on StyleGAN2 [3]. We adapt pre-trained source GANs to target domains with batch size 4 on a single NVIDIA TITAN RTX GPU. The learning rate is set as 0.002. Adam optimizer [46] is used to update trainable parameters. The adapted models are trained for 1K-3K iterations. Our approach is mainly compared with prior works on 10-shot adaptation tasks qualitatively and quantitatively.

**Baselines** We employ several baselines sharing similar targets with us to adapt source GANs to target domains with only a few available samples for comparison: TGAN [9], TGAN+ADA [12], FreezeED [7], MineGAN [10], EWC [6], CDC [14], DCL [15], RSSA [24], AdAM [26], MaskedGAN [20], and LFS-GAN [22]. We do not include BSA [8] and FSGAN [11] since they fail to produce better results than directly fine-tuned StyleGAN2 in most cases.

**Datasets** StyleGAN2 [3] models pre-trained on large-scale datasets including FFHQ [4], LSUN Church and Cars [17] are used as source models. We evaluate our approach with few-shot adaptation to various target domains, including Sketches [47], FFHQ-Babies (Babies), FFHQ-Sunglasses (Sunglasses) [4], CelebAHQ-Male and Female [22], [48], face...
Fig. 4. 10-shot image generation results on FFHQ → Sketches. Our approach adapts diverse information different from few-shot data including hairstyles, facial expressions, sunglasses, and hats, to target domains more naturally than existing methods, achieving higher quality and more diverse results. Typical comparison samples are highlighted in red boxes.

Evaluation metrics Few-shot image generation aims to generate high-quality images with great diversity using only a few available samples. Therefore, we employ FID [50], precision and recall (PR) [51], and Intra-LPIPS [14] to evaluate the quality and diversity of generation results.

FID and PR are applied to evaluate the ability of generators to reproduce real distributions using the distribution distance between generated results and real data. However, they become unstable and unreliable for datasets containing only a few samples. Therefore, we use datasets containing relatively abundant data for reliable results. The standard deviations of FID results in this paper are computed across 5 runs.

We follow prior works [14] to compute Intra-LPIPS based on LPIPS [52], which evaluates the perceptual distances [53] between images for generation diversity evaluation. We generate 1000 images with adapted models and assign them to one of the training samples with the closest perceptual distance (lowest LPIPS value). The Intra-LPIPS metric is computed with the average pairwise LPIPS within every cluster averaged over all the clusters. With exactly replicated training samples, the Intra-LPIPS metric will have a score of zero. More diverse synthesized samples correspond to larger Intra-LPIPS values. We compute Intra-LPIPS with fixed noise inputs to compare our approach with baselines fairly. The standard deviations of Intra-LPIPS results for k-shot adaptation setups in this paper are computed across k clusters (k > 1).

In addition, we conduct a user study to compare our approach with several strong baselines including CDC [14], DCL [15], RSSA [24], and AdAM [26]. We employ 40 participants to choose the best results among 4 samples synthesized.
A. 10-Shot Adaptation Setups

In this section, we follow most prior works to evaluate our approach under 10-shot adaptation setups. The proposed approach is compared with baselines through visualized samples and quantitative results, including FID, PR, Intra-LPIPS, and user study under different adaptation setups.

Qualitative evaluation We visualize the adapted samples on 10-shot FFHQ → Sketches, Amedeo’s paintings, Babies, and LSUN Church → Van Gogh houses of our approach and baselines in Fig. 4, 5, 6, and 7. For fair comparison, we exhibit results generated from fixed noise inputs for different methods. We find that TGAN and TGAN+ADA overfit seriously and tend to replicate the training samples. FreezeD and EWC add limitations to preserve parameters in the discriminator and generator, respectively. These approaches can produce some different results but still replicate training samples and cannot naturally transfer characteristics from source domains to target domains. MineGAN adds additional networks to adjust the distributions of noise inputs. MineGAN still produces some replications of the training samples and gets unnatural results. Our approach builds a better correspondence between the source and target domains than CDC and DCL. As shown in the visualized results, our approach generates diverse images and preserves detailed characteristics of source domains, achieving higher quality and greater diversity than existing approaches. For example, our approach inherits the occlusion of hats and sunglasses from the source domain more naturally. Moreover, hairstyles and facial expressions are better preserved as well. In Fig. 7, our approach preserves more detailed house structures while adapting to the target domain. RSSA is designed to preserve the global image structures learned from source domains. As a result, it may be inappropriate for adaptation to abstract target domains like artists’ paintings. Taking Fig. 5 as an example, RSSA tends to produce wider faces while our approach achieves narrower...
Fig. 6. 10-shot image generation results on FFHQ → Babies. Our approach produces more realistic results with fewer blurs and artifacts than baselines. Typical comparison samples are highlighted in red boxes.

faces like training data when adapting FFHQ to Amedeo’s paintings. Similarly, our approach achieves better learning of target distributions when adapting LSUN Church to Van Gogh houses. In addition, our approach preserves facial expressions better and produces more realistic results with fewer blurs and artifacts in FFHQ → Sketches and Babies. AdAM focuses on unrelated source/target domain adaptation like FFHQ → AFHQ Cat [54]. Our approach produces more diverse and realistic results under the employed adaptation setups.

We add the comparison between our approach and LFS-GAN and MaskedGAN in Fig. 8. LFS-GAN proposes a lifelong few-shot image generation method but fails to produce compelling results like our approach. More detailed comparison with LFS-GAN is provided in Appendix. MaskedGAN applies patch-level spatial masks and spectral masks at the image level. It produces blurred samples lacking details with limited training data. Combining MaskedGAN with CDC loss achieves better preservation of source information but still gets unnatural adaptation with too many blurs and artifacts.

More results of our approach under different adaptation setups, including FFHQ → Otto’s paintings, LSUN Cars → Wrecked cars, and LSUN Church → Haunted houses, are shown in Fig. 9. Our approach also achieves pleasing visual effects while maintaining diversity under these adaptation setups. Additional adapted samples of different target domains using the same source domain FFHQ are shown in Fig. 10.

Quantitative evaluation To evaluate the capability of our approach to model real distributions, we use datasets containing relatively abundant images for stable FID and PR evaluation, including the original Sketches, FFHQ-Babies, and FFHQ-Sunglasses datasets, which roughly contain 300, 2500, and 2700 images, respectively. Our approach outperforms all the baselines remarkably under 10-shot adaptation setups as shown in Tables I and II, demonstrating its strong ability to reproduce target distributions using limited data.

We evaluate generation diversity under several 10-shot adaptation setups using the source domain FFHQ and report the Intra-LPIPS results in Table III. Our approach outer-
forms baselines on almost all the benchmarks in terms of Intra-LPIPS when adapted from the source domain FFHQ. To further prove the robustness of our approach, we compare it with baselines on additional 10-shot adaptation tasks: LSUN Church → Haunted houses and Van Gogh houses, and LSUN Cars → Wrecked cars as shown in Table IV and find improved results as well. With masked discrimination, our approach guides the few-shot adapted GANs to generate more diverse images different from training samples. We provide the results of the user study in Table V. Our approach relieves overfitting and produces more diverse results with fewer blurs and artifacts, obtaining the highest support rates under all the employed 10-shot adapta-
Fig. 8. Visualized results of LFS-GAN and MaskedGAN on 10-shot FFHQ → Sketches, Babies, and Amedeo’s paintings.

Fig. 9. Additional 10-shot image generation results of our approach under different adaptation setups including FFHQ → Otto’s paintings, LSUN Cars → Wrecked cars, and LSUN Church → Haunted houses.

TABLE III

| Intra-LPIPS (↑) Results of 10-Shot Image Generation Tasks Adapted From the Source Domain FFHQ |
|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| Approaches | FFHQ → Sketches | FFHQ → Babies | FFHQ → Sunglasses | FFHQ → Otto’s paintings | FFHQ → Raphael’s paintings | FFHQ → Amedeo’s paintings |
|------------|----------------|---------------|------------------|-------------------------|---------------------------|---------------------------|
| TGAN [9]   | 0.394 ± 0.023 | 0.510 ± 0.026 | 0.556 ± 0.023    | 0.594 ± 0.023           | 0.533 ± 0.023             | 0.548 ± 0.026             |
| TGAN+ADA [12] | 0.427 ± 0.022 | 0.546 ± 0.033 | 0.571 ± 0.034    | 0.625 ± 0.028           | 0.546 ± 0.037             | 0.560 ± 0.019             |
| FreezeD [7] | 0.406 ± 0.017 | 0.535 ± 0.021 | 0.558 ± 0.024    | 0.629 ± 0.023           | 0.537 ± 0.026             | 0.558 ± 0.019             |
| MineGAN [10] | 0.407 ± 0.020 | 0.514 ± 0.034 | 0.570 ± 0.020    | 0.625 ± 0.030           | 0.559 ± 0.031             | 0.586 ± 0.041             |
| EWC [6]     | 0.430 ± 0.018 | 0.560 ± 0.019 | 0.550 ± 0.014    | 0.611 ± 0.025           | 0.541 ± 0.023             | 0.579 ± 0.035             |
| CDC [14]    | 0.454 ± 0.017 | 0.583 ± 0.014 | 0.581 ± 0.011    | 0.638 ± 0.023           | 0.564 ± 0.010             | 0.620 ± 0.029             |
| DCL [15]    | 0.461 ± 0.021 | 0.579 ± 0.018 | 0.574 ± 0.007    | 0.617 ± 0.033           | 0.558 ± 0.033             | 0.616 ± 0.043             |
| RSSA [24]   | 0.477 ± 0.016 | 0.583 ± 0.011 | 0.569 ± 0.021    | 0.620 ± 0.023           | 0.562 ± 0.024             | 0.611 ± 0.022             |
| AdAM [26]   | 0.455 ± 0.021 | 0.582 ± 0.023 | 0.575 ± 0.017    | 0.623 ± 0.025           | 0.567 ± 0.033             | 0.576 ± 0.045             |
| Ours        | 0.505 ± 0.020 | 0.595 ± 0.006 | 0.593 ± 0.014    | 0.638 ± 0.024           | 0.581 ± 0.012             | 0.628 ± 0.024             |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
B. 5-Shot and 1-Shot Adaptation Setups

We add experiments using 5-shot and 1-shot target datasets. We provide visualized samples of our approach and baselines in Fig. 11. Given 5 training samples, our approach and RSSA produce diverse samples while TGAN and TGAN+ADA overfit and synthesize samples very similar to training data. Our approach adapts source samples to target domains more naturally than RSSA and preserves diverse features of source samples better. In 1-shot adaptation tasks, our approach still maintains considerable diversity. However, RSSA produces faces as same as training samples in 1-shot adaptation from FFHQ and fails to preserve detailed house structures in 1-shot LSUN Church → Haunted houses. Moreover, our approach achieves greater generation diversity than baselines in terms of Intra-LPIPS, as shown in Table VI.

C. Ablations

**Ablations of each component** Compared with existing few-shot image generation methods, our approach proposes two new ideas, random masks applied to features in the target discriminator and discriminator CDC loss. As shown in Fig 12, we provide visualized ablations of our approach using 10-shot FFHQ → Sketches as an example. Masked discrimination encourages the target discriminator to judge more diverse images as realistic and guides the adapted GAN to learn the common features of limited training samples. Absence of random masks makes it hard for the adapted GAN to preserve characteristics like hairstyles and the occlusion of hats and sunglasses, which are obviously different from the training samples. Discriminator CDC loss helps strengthen global image discrimination and preserve information learned from the source domain. Without it, the adapted GAN produces results containing unnatural blurs and artifacts, leading to degraded image quality. Moreover, we provide a quantitative ablation analysis in terms of Intra-LPIPS and FID in Tables VII and VIII. It can be seen that both ideas contribute to improving generation diversity and learning of...
Fig. 11. 5-shot and 1-shot image generation results of our approach and baselines. Our approach maintains considerable diversity with extremely limited data and produces more realistic results with fewer blurs and artifacts than baselines.

TABLE V
USER STUDY CONDUCTED BY 40 PARTICIPANTS UNDER 10-SHOT ADAPTATION SETUPS. WE ASK USERS TO CONSIDER BOTH QUALITY AND DIVERSITY GIVEN SOURCE AND TRAINING SAMPLES AS REFERENCE. OUR APPROACH ACHIEVES HIGHER SUPPORT RATES THAN BASELINES.

| Approaches | FFHQ → Sketches | FFHQ → Babies | FFHQ → Sunglasses | FFHQ → Amedeo’s paintings | FFHQ → Otto’s paintings | FFHQ → Raphael’s paintings | LSUN Church → Haunted houses | LSUN Church → Van Gogh Houses | LSUN Cars → Wrecked Cars |
|------------|-----------------|---------------|-------------------|---------------------------|------------------------|---------------------------|----------------------------|----------------------------|--------------------------|
| CDC [14]   | 10.00%          | 6.25%         | 18.75%            | 21.88%                    | 9.33%                  | 16.88%                    | 19.38%                     | 23.75%                     | 16.83%                   |
| DCL [15]   | 2.50%           | 8.75%         | 11.24%            | 3.75%                     | 12.50%                 | 8.75%                     | 6.88%                      | 5.63%                      | 6.25%                    |
| RSSA [24]  | 28.13%          | 26.13%        | 8.33%             | 1.88%                     | 25.00%                 | 31.88%                    | 17.50%                     | 5.63%                      | 30.00%                   |
| AdAM [26]  | 8.13%           | 10.63%        | 6.25%             | 8.75%                     | 14.38%                 | 1.25%                     | 16.88%                     | 4.38%                      | 12.50%                   |
| Ours       | 51.25%          | 48.25%        | 54.34%            | 63.75%                    | 38.75%                 | 43.13%                    | 36.88%                     | 61.63%                     | 34.38%                   |

TABLE VI
INTRA-LPIPS (↑) RESULTS OF 5-SHOT AND 1-SHOT IMAGE GENERATION TASKS. THE STANDARD DEVIATIONS OF 1-SHOT ADAPTATION TASKS ARE COMPUTED ACROSS 1000 SAMPLES SYNTHESIZED FROM FIXED NOISE INPUTS.

| Datasets   | FFHQ → Amedeo’s paintings | FFHQ → Babies | LSUN Church → Haunted houses |
|------------|---------------------------|---------------|-----------------------------|
| K-shot     | 5-shot        | 1-shot        | 5-shot          | 1-shot          | 5-shot         |
| TGAN [9]   | 0.521 ± 0.010 | 0.343 ± 0.011 | 0.526 ± 0.008 | 0.342 ± 0.011 | 0.555 ± 0.009 | 0.369 ± 0.048 |
| TGAN+ADA [12] | 0.565 ± 0.016 | 0.385 ± 0.113 | 0.539 ± 0.007 | 0.393 ± 0.094 | 0.555 ± 0.031 | 0.427 ± 0.048 |
| RSSA [24]  | 0.589 ± 0.033 | 0.507 ± 0.049 | 0.598 ± 0.019 | 0.537 ± 0.049 | 0.625 ± 0.016 | 0.612 ± 0.038 |
| Ours       | 0.615 ± 0.015 | 0.628 ± 0.032 | 0.607 ± 0.015 | 0.560 ± 0.042 | 0.642 ± 0.011 | 0.635 ± 0.040 |

target distributions, while masked discrimination contributes more to the overall improvement of evaluation metrics. Our full approach performs well on visual effects and quantitative evaluation, benefiting from the combination of these two ideas.
Ablations of mask size We apply random masks to features extracted by the target discriminator, encouraging adapted GANs to generate more diverse images. As shown in Fig. 13, we use different sizes of masks on 10-shot FFHQ → Amedeo’s paintings for ablation analysis. With larger masks, the adapted GANs tend to preserve more information learned from the source domain and produce images that are different from the training samples. For example, faces occupy the maximum ratio of the whole image in results generated from the adapted GAN using 7/8 masked features. In contrast, most of the training samples only occupy a small part. With too small masks, it is hard for adapted GANs to naturally transfer characteristics of source images, such as facial expressions, hairstyles, and sunglasses, to target domains. We recommend mask ratios ranging from 1/2 to 3/4. The weight coefficient of CDC loss ($\lambda$) is set as 2500 in the ablations of mask size.

Ablations of $\lambda$ In Fig. 14, we visualize 10-shot image generation results on FFHQ → Amedeo’s paintings using different $\lambda$ ranging from 0 to 7500. Unnatural blurs can be found in generated images using too small $\lambda$. With larger $\lambda$, the adapted GANs focus more on keeping characteristics learned from the source domain. Too large $\lambda$ can prevent adapted GANs from learning from the target domain and lead to unrealistic results. We recommend $\lambda$ ranging from 1000 to 5000. We use the same weight coefficient for cross-domain consistency loss applied to the generator and discriminator and achieve compelling results. Diverse parameter combinations for the generator and discriminator can be tried for better results. We randomly mask 3/4 features extracted by the target discriminator in ablations of $\lambda$.

Ablations of the masked layer choice We propose masked discrimination to apply random masks to the feature maps output by discriminator block $4^2$ as shown in Fig. 2. We tried to apply random masks to the outputs of different layers of the discriminator and provide qualitative and quantitative evaluation as follows. As shown in Table IX, the approach used in the paper consistently, which masks the output of different $\lambda$. The results in Table IX show that the approach consistently improves image quality with different $\lambda$. The table includes comparisons with results obtained using different mask sizes and $\lambda$ values.

**Table VIII**

| Adaptation Setups   | Ours w/o Random Masks | Ours w/o $L_{dist}(D_s, D_t)$ | Full Approach |
|---------------------|-----------------------|-------------------------------|---------------|
| FFHQ→Sketches       | 38.83 ± 0.01          | 34.16 ± 0.02                  | 28.93 ± 0.01  |
| FFHQ→Raphael’s paintings | 52.91 ± 0.01          | 43.63 ± 0.02                  | 36.39 ± 0.01  |
| FFHQ→Sunglasses     | 36.89 ± 0.01          | 33.22 ± 0.03                  | 26.96 ± 0.01  |
Fig. 14. Visualized ablations of $\lambda$, the weight coefficient of cross-domain consistency loss applied to the generator and discriminator, on 10-shot FFHQ $\rightarrow$ Amedeo’s paintings.

![Visualization of ablations for different values of $\lambda$.](image)

Fig. 15. Visualized ablations of the masked layer choice in the discriminator on 10-shot FFHQ $\rightarrow$ Babies.

![Visualization of ablations for different masked layers.](image)

TABLE IX

| Masked Layer | FFHQ $\rightarrow$ Babies | FFHQ $\rightarrow$ Sunglasses |
|--------------|---------------------------|------------------------------|
| discriminator block $256^2$ | $0.582 \pm 0.020$ | $0.581 \pm 0.012$ |
| discriminator block $128^2$ | $0.572 \pm 0.014$ | $0.576 \pm 0.016$ |
| discriminator block $64^2$ | $0.575 \pm 0.012$ | $0.572 \pm 0.014$ |
| discriminator block $32^2$ | $0.568 \pm 0.017$ | $0.560 \pm 0.009$ |
| discriminator block $16^2$ | $0.590 \pm 0.022$ | $0.574 \pm 0.019$ |
| discriminator block $8^2$ | $0.590 \pm 0.018$ | $0.576 \pm 0.011$ |
| discriminator block $4^2$ | $\mathbf{0.595 \pm 0.006}$ | $\mathbf{0.593 \pm 0.014}$ |
| first linear layer | $0.590 \pm 0.013$ | $0.584 \pm 0.017$ |

Fig. 16. Visualized samples of applying feature-level masks (our approach) and image-level masks to the discriminator on 10-shot FFHQ $\rightarrow$ Amedeo’s paintings.

![Visualization of samples with feature-level and image-level masks.](image)

D. Feature-Level Masks v.s. Image-Level Masks

We apply random masks to the output features of an intermediate layer in the discriminator. It can be seen as a feature-level augmentation approach. Here we provide qualitative comparison between image-level and feature-level masks using 10-shot FFHQ $\rightarrow$ Amedeo’s paintings as an example in Fig. 16. When applying image-level random masks, we get low-quality and blurred results sharing similar structures like eyes and teeth. In addition, image-level masks lead to lower Intra-LPIPS results than our approach as shown in Table IX.

Compared with feature-level masks, image-level masks make it harder for target discriminators to judge image quality. A masked high-quality image can be similar to a masked blurry image from the view of discriminators. Masked images also affect the computation of discriminator CDC loss, leading to degraded quality and diversity. The proposed feature-level masked discrimination guides target discriminators to focus on the common features of training samples, which have greater possibilities to be preserved with random masks. In this way, target discriminators are less vulnerable to overfitting and guide adapted models to produce more diverse results. Our
TABLE X
INTRA-LPIPS (↑) RESULTS COMPARISON BETWEEN APPLYING IMAGE-LEVEL AND FEATURE-LEVEL RANDOM MASKS

| Methods | FFHQ → Babies | FFHQ → Amedeo’s paintings | LSUN Church → Haunted houses |
|---------|---------------|---------------------------|-----------------------------|
| Image   | 0.581 ± 0.033 | 0.602 ± 0.027             | 0.605 ± 0.009               |
| Feature (Ours) | 0.595 ± 0.006 | 0.628 ± 0.024                | 0.632 ± 0.020               |

Fig. 17. 10-shot image generation results of our approach on 1024 × 1024 FFHQ → Babies (upper) and Sunglasses (bottom).

TABLE XI
INTRA-LPIPS (↑) RESULTS ON 10-SHOT FFHQ → BABIES AND SUNGLASSES (1024 × 1024)

| Methods | FFHQ → Babies | FFHQ → Sunglasses |
|---------|---------------|-------------------|
| CDC [14] | 0.597 ± 0.029 | 0.606 ± 0.017     |
| Ours    | 0.610 ± 0.026 | 0.617 ± 0.014     |

E. Higher Resolution Experiments
Most prior researches work on the resolution of 128 × 128 and 256 × 256. We mainly conduct experiments with the resolution of 256 × 256 for comparison with baselines, except for LSUN Cars to Wrecked cars, which has the resolution of 512 × 512. We add experiments on 10-shot FFHQ → Babies and Sunglasses with the resolution of 1024 × 1024. As shown in Fig. 17, our approach can be applied to high-resolution datasets and synthesize high-quality and diverse results. In addition, we compare our approach with CDC [14], which performs better than other baselines on 10-shot FFHQ → Babies and Sunglasses (256 × 256) in terms of Intra-LPIPS in Table XI. It can be seen that our approach outperforms CDC on the resolution of 1024 × 1024 as well.

F. BigGAN-Based Masked Discrimination
We follow prior works to implement our approach with the powerful StyleGAN2 [3] for high-quality results and fair comparison. We further adapt our approach to UNetGAN [55], which applies a UNet-based discriminator to BigGAN [2]. The proposed masked discrimination is compatible with various GANs while the CDC loss [14] is designed for StyleGAN2. Therefore, we only apply masked discrimination to UNetGAN and compare it with the TGAN [9] baseline, both of which are trained for 5000 iterations. As shown in Fig. 18, our approach significantly improves the generation quality with fewer blurs and preserves more details like teeth, hairstyles, and facial expressions. Our approach also achieves greater diversity in terms of Intra-LPIPS, as shown in Table XII.

G. Computational Cost
Training Time Cost We list the computational cost of different few-shot image generation approaches in Table XIII. We conduct all the experiments under 10-shot adaptation setups for 3000 iterations with the same batch size 4 on a single NVIDIA TITAN RTX GPU. MineGAN [10] needs a two-stage training strategy. In the first stage, the generator is fixed and the miner is optimized with the discriminator. Then the miner, generator, and discriminator are optimized together in the second stage. RSSA [24] projects real samples into the latent space before adaptation to target domains. AdAM [26] identifies important kernels for domain adaptation first. Therefore, we provide the time cost of them in “X+Y” format, measuring two stages separately. The time cost of our approach is comparable to most previous methods and lower than several recent works including DCL [15], RSSA [24], and AdAM [26]. We provide the training time cost with resolution 256 × 256 and 1024 × 1024. AdAM needs too large GPU memory and cannot be conducted on a single GPU of 48 memory even with batch size 1 for resolution 1024 × 1024.

Inference Time Cost Most baselines and our approach fine-tune pre-trained models without modification on network structures or inference processes. As a result, they share the same time cost for inference as StyleGAN2. MineGAN [10] adds additional MLP networks to optimize the distributions of latent codes. RSSA [24] compresses the latent space for target domain optimization. As a result, they cost much more time than other approaches. We provide the inference time cost of resolution 256 × 256 and 1024 × 1024 with the same batch size 1 on a single NVIDIA TITAN RTX GPU in Table XIV.
and wrinkles for babies, as shown in Fig. 6. It remains a
serve features inappropriate for target domains like beards
quality and greater diversity than prior works.
learned from source domains. Our approach is simple and
achieve higher image quality by preserving more information
on regulating the discriminator, which easily overfits and
ation task using extremely limited data. Our approach focuses
while our approach has achieved compelling results, it is
not without limitations. For example, our approach may pre-
sure features inappropriate for target domains like beards
and wrinkles for babies, as shown in Fig. 6. It remains a
challenge for adapted GANs to discriminate what kind of
features should be preserved or modified with only a few
available real samples. Nevertheless, our approach has made
significant progress in few-shot image generation tasks and
achieved high-quality and diverse results. We hope that our
work will be a solid basis for better methods in the future.

V. CONCLUSION AND LIMITATIONS

This paper explores the challenging few-shot image gener-
task using extremely limited data. Our approach focuses
on regulating the discriminator, which easily overfits and
guides the generator to replicate training samples in few-shot
image generation tasks. We propose the first feature-level
augmentation method for generative tasks, named masked
discrimination. It applies random masks to features extracted
by the discriminator to encourage adapted GANs to learn
the common features of limited training samples and produce
more diverse images. Besides, we design the discriminator
CDC loss to strengthen global image discrimination and
achieve higher image quality by preserving more information
learned from source domains. Our approach is simple and
proven effective qualitatively and quantitatively, with higher
quality and greater diversity than prior works.

While our approach has achieved compelling results, it is
not without limitations. For example, our approach may pre-
sure features inappropriate for target domains like beards
and wrinkles for babies, as shown in Fig. 6. It remains a
challenge for adapted GANs to discriminate what kind of
features should be preserved or modified with only a few
available real samples. Nevertheless, our approach has made
significant progress in few-shot image generation tasks and
achieved high-quality and diverse results. We hope that our
work will be a solid basis for better methods in the future.

REFERENCES

[1] I. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural
Inf. Process. Syst., vol. 27, 2014, pp. 1–7.
[2] A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training
for high fidelity natural image synthesis,” in Proc. Int. Conf. Learn. Represent., 2019, pp. 1–35.
[3] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila,
“Analyzing and improving the image quality of StyleGAN,” in Proc.
IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 8110–8119.
[4] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture
for generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4401–4410.
[5] Q. Feng, C. Guo, F. Benitez-Quiroz, and A. Martinez, “When do GANs
replicate? On the choice of dataset size,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 6681–6690.
[6] Y. Li, R. Zhang, J. Lu, and E. Shechtman, “Few-shot image generation
with elastic weight consolidation,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 15885–15896.
[7] S. Mo, M. Cho, and J. Shin, “Freeze the discriminator: A simple baseline
for fine-tuning gans,” in Proc. CVPR AI Content Creation Workshop, 2020, pp. 1–13.
[8] A. Noguchi and T. Harada, “Image generation from small datasets via
batch statistics adaptation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 9332–9341.
[9] E. Robb, W.-S. Chu, A. Kumar, and J.-B. Huang, “Few-shot adaptation
of generative adversarial networks,” 2020, arXiv:2010.11943.
[10] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila,
“Training generative adversarial networks with limited data,” in Proc.
34th Int. Conf. Neural Inf. Process. Syst., 2020, pp. 12104–12114.
[11] X. Zhao, Z. Liu, J. Lin, Y.-Z. Zhu, and S. Han, “Differentiable augmenta-
tion for data-efficient GAN training,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 7559–7570.
[12] Y. Wang, A. Gonzalez-Garcia, D. Berga, L. Herranz, F. S. Khan, and
J. van de Weijer, “MineGAN: Effective knowledge transfer from GANs
to target domains with few images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9332–9341.
[13] S. Zhao, Z. Liu, J. Lin, Y.-Z. Zhu, and S. Han, “Differentiable augmenta-
tion for data-efficient GAN training,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 7559–7570.
[14] U. Ojha et al., “Few-shot image generation via cross-domain corre-
spondence,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2021, pp. 10743–10752.
[15] Y. Zhao, H. Ding, H. Huang, and N.-M. Cheung, “A closer look at few-
shot image generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 9130–9140.
[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet:
A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Miami, FL, USA, Jun. 2009, pp. 248–255.
[17] F. Yu, A. Self, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, “LSUN:
Construction of a large-scale image dataset using deep learning with
humans in the loop,” 2015, arXiv:1506.03365.
[18] N. Tran, V. Tran, N. Nguyen, T. Nguyen, and N. Cheung, “On data
augmentation for GAN training,” IEEE Trans. Image Process., vol. 30, pp. 1882–1897, 2021.
[19] Z. Zhao, Z. Zhang, T. Chen, S. Singh, and H. Zhang, “Image augmenta-
tions for GAN training,” 2020, arXiv:2006.02595.
[20] J. Huang et al., “Masked generative adversarial networks are data-
efficient generation learners,” in Proc. Adv. Neural Inf. Process. Syst., vol. 35, 2022, pp. 2154–2167.
[21] A. Ly, M. Marsman, J. Verhagen, R. P. P. Grasman, and J.-E. Wagenmakers,
“A tutorial on Fisher information,” J. Math. Psychol., vol. 80, pp. 40–55, Oct. 2017.
[22] J. Seo, J.-S. Kang, and G.-M. Park, “LFS-GAN: Lifelong few-shot
image generation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2023, pp. 11356–11366.
[23] Z. Zhu, H. Ma, J. Chen, and J. Yuan, “Few-shot image generation with
diffusion models,” 2022, arXiv:2211.03264.
[24] J. Xiao, L. Li, C. Wang, Z.-J. Zha, and Q. Huang, “Few shot generative
model adaption via relaxed spatial structural alignment,” in Proc.
IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 11194–11203.
[25] R. Abdal, Y. Qin, and P. Wonka, “Image2StyleGAN++: How to edit
the embedded images?” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 8293–8302.
[26] Y. Zhao, K. Chandrasegaran, M. Abdollahzadeh, and N.-M. Cheung,
“Few-shot image generation via adaptation-aware kernel modulation,” in Proc. Adv. Neural Inf. Process. Syst., 2022, pp. 19427–19440.
[27] Y. Zhao et al., “Exploring incompatible knowledge transfer in few-
shot image generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023, pp. 7380–7391.
[28] C. Yang et al., “One-shot generative domain adaptation,” in Proc.
IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2023, pp. 7733–7742.
[29] P. Zhu, R. Abdal, J. Fieniami, and P. Wonka, “Mind the gap: Domain
gap control for single shot domain adaptation for generative adversarial
networks,” in Proc. Int. Conf. Learn. Represent., 2021, pp. 1–20.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
[30] G. Kwon and J. C. Ye, “One-shot adaptation of GAN in just one CLIP,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 10, pp. 12179–12191, Oct. 2023.

[31] Z. Zhang, Y. Liu, C. Han, T. Guo, T. Yao, and T. Mei, “Generalized one-shot domain adaption of generative adversarial networks,” in Proc. Adv. Neural Inf. Process. Syst., 2022, pp. 13718–13730.

[32] Y. Zhang, M. Yao, Y. Wei, Z. Ji, J. Bai, and W. Zuo, “Towards diverse and faithful one-shot adaption of generative adversarial networks,” in Proc. Adv. Neural Inf. Process. Syst., 2022, pp. 37297–37308.

[33] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Mach. Learn., vol. 139, 2021, pp. 8748–8763.

[34] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 1125–1134.

[35] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. ICCV, Oct. 2017, pp. 2223–2232.

[36] J.-Y. Zhu et al., “Toward multimodal image-to-image translation,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1–12.

[37] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, “Diverse image-to-image translation via disentangled representations,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 35–51.

[38] L. Ma, X. Jia, S. Georgoulis, T. Tuytelaars, and L. Van Gool, “Exemplar guided unsupervised image-to-image translation with semantic consistency,” in Proc. Int. Conf. Learn. Represent., 2019, pp. 1–17.

[39] E. Richardson et al., “Encoding in style: A styleGAN encoder for image-to-image translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 2287–2296.

[40] Y. Pang, J. Lin, T. Qin, and Z. Chen, “Image-to-image translation: Methods and applications,” IEEE Trans. Multimedia, vol. 24, pp. 3859–3881, 2022.

[41] K. Saito, K. Saenko, and M.-Y. Liu, “COCO-FUNIT: Few-shot unsupervised image translation with a content conditioned style encoder,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 382–398.

[42] Y. Wang, S. Khan, A. Gonzalez-Garcia, J. V. D. Weijer, and F. S. Khan, “Semi-supervised learning for few-shot image-to-image translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2020, pp. 4453–4462.

[43] G. Kang, X. Dong, L. Zheng, and Y. Yang, “PatchShuffle regularization,” 2017, arXiv:1707.07103.

[44] D. Warde-Farley, I. J. Goodfellow, A. Courville, and Y. Bengio, “An empirical analysis of dropout in piecewise linear networks,” 2013, arXiv:1312.6197.

[45] V. K. Kurmi, V. K. Subramanian, and V. P. Namboodiri, “Exploring dropout discriminator for domain adaptation,” Neurocomputing, vol. 457, pp. 168–181, Oct. 2021.

[46] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

[47] X. Wang and X. Tang, “Face photo-sketch synthesis and recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 11, pp. 1955–1967, Nov. 2008.

[48] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in Proc. Int. Conf. Learn. Represent., 2017, pp. 1–26.

[49] J. Yaniv, Y. Newman, and A. Shamir, “The face of art: Landmark detection and geometric style in portraits,” ACM Trans. Graph., vol. 38, no. 4, pp. 1–15, Aug. 2019.

[50] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1–12.

[51] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, “Assessing generative models via precision and recall,” in Proc. Adv. Neural Inf. Process. Syst., vol. 31, 2018, pp. 1–10.

[52] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 586–595.

Jingyuan Zhu received the B.S. degree in aerospace engineering from Tsinghua University in 2020, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering.

Huimin Ma (Senior Member, IEEE) received the M.S. and Ph.D. degrees in mechanical and electronic engineering from Beijing Institute of Technology, Beijing, China, in 1998 and 2001, respectively. She is currently a Professor with the School of Computer and Communication Engineering, University of Science and Technology Beijing. Her research and teaching interests include 3D object recognition and tracking, system modeling and simulation, and psychological base of image cognition.

Jiansheng Chen (Senior Member, IEEE) received the M.S. degree from the Department of Computer Science and Technology, Tsinghua University, in 2002, and the Ph.D. degree in computer science and technology from The Chinese University of Hong Kong in 2006. He is currently a Professor with the School of Computer and Communication Engineering, University of Science and Technology Beijing. His research and teaching interests include computer vision and machine learning.

Jian Yuan received the Ph.D. degree in communication and electronic systems from the University of Electronic Science and Technology of China. He is currently a Professor with the Department of Electronic Engineering, Tsinghua University. His research and teaching interests include complex network theory and technologies.