Few-shot Image Generation via Cross-domain Correspondence

Utkarsh Ojha[1,2] Yijun Li[1] Jingwan Lu[1] Alexei A. Efros[1,3]
Yong Jae Lee[2] Eli Shechtman[1] Richard Zhang[1]

[1]Adobe Research [2]UC Davis [3]UC Berkeley

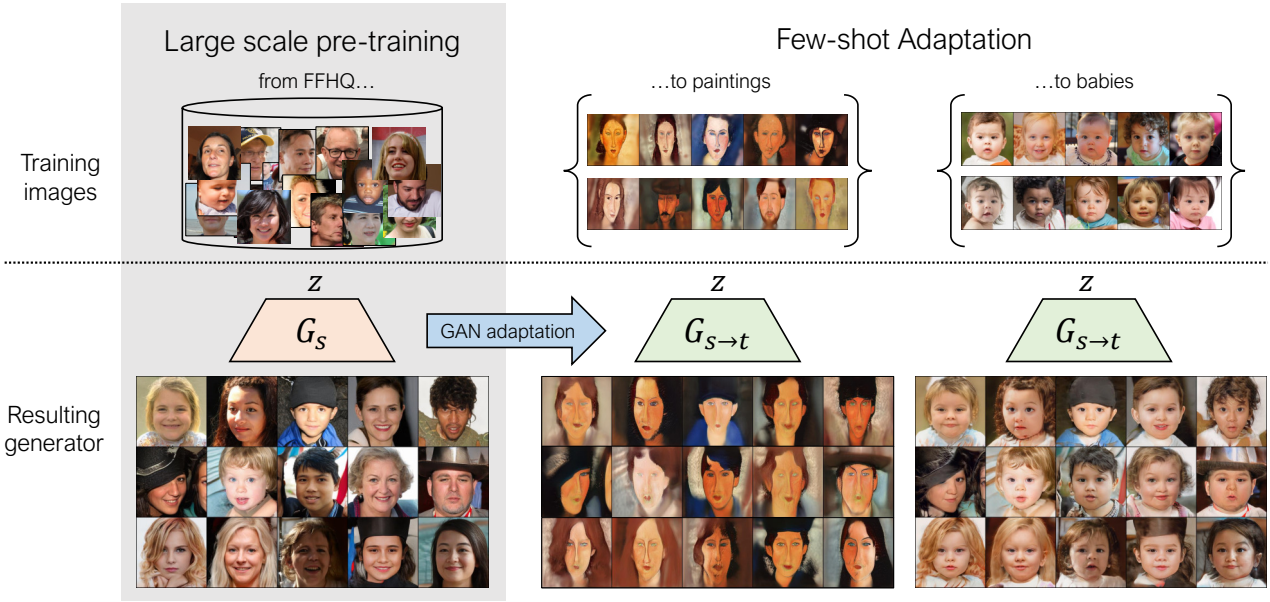

Figure 1: Given a model trained on a large source dataset (G_s), we propose to adapt it to arbitrary image domains, so that the resulting model ($G_{s\to t}$) captures these target distributions using extremely few training samples. In the process, our method discovers a one-to-one relation between the distributions, where noise vectors map to corresponding images in the source and target. Consequently, one can imagine how a natural face would look if Amedeo Modigliani had painted it, or how the face would look in its baby form.

Abstract

Training generative models, such as GANs, on a target domain containing limited examples (e.g., 10) can easily result in overfitting. In this work, we seek to utilize a large source domain for pretraining and transfer the diversity information from source to target. We propose to preserve the relative similarities and differences between instances in the source via a novel cross-domain distance consistency loss. To further reduce overfitting, we present an anchor-based strategy to encourage different levels of realism over different regions in the latent space. With extensive results in both photorealistic and non-photorealistic domains, we demonstrate qualitatively and quantitatively that our few-shot model automatically discovers correspondences between source and target domains and generates more diverse and realistic images than previous methods.

1. Introduction

Consider 10 portrait paintings by the incomparable Amedeo Modigliani [34], shown in Fig. 1 (middle). Given only these 10 paintings, would it be possible to train a model which can generate infinitely many paintings in the style of Modigliani? Unfortunately, contemporary generative models [11, 12, 13, 27, 3] require thousands of images to train properly, not 10. This problem is of practical importance, since many such domains of interest have a very limited collection of images (e.g., there are just 10 examples per artist in the Artistic-Faces dataset [34]).

Transfer learning serves as an alternative to training from scratch and has been explored in the context of generative

adversarial networks (GANs) to address the limited data regime. The key idea is to start with a source model, pretrained on a large dataset, and adapt it to a target domain with limited data by either making only small changes to the network parameters to preserve as much information as possible [32, 21, 30, 19, 15], or by synthetically increasing the training data via data augmentation [37, 11]. Most of these methods, however, are designed for scenarios with more than 100 training images. When the number of available images is lowered to just a few [15], results often overfit to the training samples or are of poor quality.

In this work, we explore transferring a different kind of information from the source domain, namely *how images relate to each other*, to address the limited data setting. Intuitively, if the model can preserve the relative similarities and differences between instances in the source domain, then it has the chance to inherit the diversity in the source domain while adapting to the target domain. To capture this notion, we introduce a novel cross-domain distance consistency loss, which enforces similarity in the distribution of pairwise distances of generated samples before and after adaptation. Unlike domain adaptation approaches like image-to-image translation, here we are adapting models, not images.

Interesting properties emerge when enforcing this structure-level alignment between the two domains. Specifically, when the source and target domains are related (e.g., faces and caricatures), our approach automatically *discovers* a one-to-one correspondence between them and is able to more faithfully model the true target distribution in terms of both diversity and image realism, as shown in Fig. 1. When the two domains are unrelated (e.g., cars and caricatures), our approach is unable to model the target distribution but still discovers interesting part-level correspondences to generate diverse samples.

Since the few training samples only form a small subset of the target distribution we seek to approximate, we find it necessary to enforce realism in two different ways, to not inordinately penalize the diversity among the generated images. We apply an image-level adversarial loss on the synthesized images which should map to one of the real samples. For all other synthesized images, we only enforce a patch-level adversarial loss. In this way, only a small subset of our generated samples need to look like one of the few-shot training images, while the rest are only forced to capture their patch-level texture.

Contributions. Our main contribution is a novel GAN adaptation framework, which enforces cross-domain correspondence for few-shot image generation. Through extensive qualitative and quantitative results, we demonstrate that our model automatically discovers correspondences between related source and target domains to generate diverse and realistic images.

2. Related work

Few-shot learning. Representative few-shot classification approaches include learning a feature similarity function between the query and support examples [28, 24] and learning to learn how to adapt a base-learner to a new task [6, 20].

Few-shot image generation aims instead to hallucinate new and diverse examples while preventing overfitting to the few training images. Existing work mainly follows an adaptation pipeline, in which a base model is pretrained on a large source domain and then adapted to a smaller target domain. They either embed a small number of new parameters into the source model [21, 30] or directly update the source model parameters, using different forms of regularization [19, 15]. Others employ data augmentation to reduce overfitting [37, 11] but are less effective under the extreme few-shot setting (e.g., 10 images). In contrast to prior work, we regularize the adaptation of the source model by transferring how images relate to each other in the source domain to the target domain, leading to plausible generation results, even with very few examples.

Domain translation. Translating images from the source domain is an alternative approach for generating more target domain data. However, such methods [9, 38, 39] require a large amount of training data for both source and target domains and are not suitable for the few-shot scenario. Recent work [16, 31, 23] has begun to address this issue via learning to separate the content and style factor, but requires large amount of labeled data (class or style labels). In our case, we assume access to a large amount of unlabeled data in the source domain and focus on adapting the source model to the target domain for unconditional image generation.

Distance preservation. To alleviate mode collapse in GANs, DistanceGAN [2] proposes to preserve the distances between input pairs in the corresponding generated output pairs. A similar scheme has been employed for both unconditional [25, 17] and conditional [18, 33] generation tasks to increase diversity in the generations. In our work, we aim to inherit the learned diversity from the source model to the target model and achieve this via our novel cross-domain distance consistency loss.

3. Approach

We are given a source generator G_s, trained on a large source dataset $\mathcal{D}_s$, which maps noise vectors $z \sim p_z(z) \subset \mathcal{Z}$, drawn from a simple distribution in a low-dimensional space, into images x. We aim to learn an adapted generator $G_{s \to t}$ by initializing the weights to the source generator and fitting it to a small target dataset $\mathcal{D}_t$.

A naive translation can be obtained simply by using a

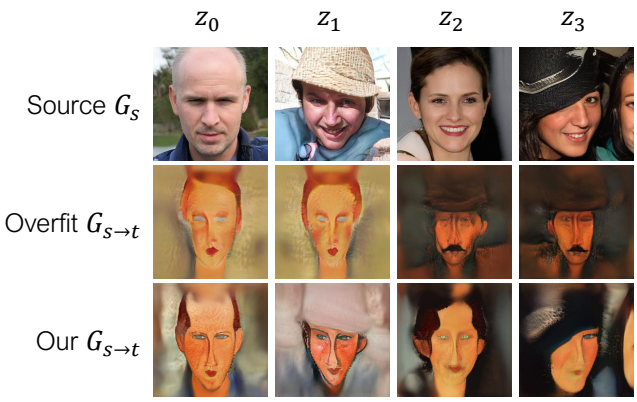

Figure 2: Adapting a source model (FFHQ) to 10 paintings using only the adversarial loss [32] results in overfitting, which leads to a loss in correspondence between source and target images. Our adaptation method preserves this property in a much better way, where the same noise maps to corresponding images in source/target.

GAN training procedure, with a learned discriminator D. With the non-saturating GAN objective, this corresponds to solving:

$$\begin{aligned} \mathcal{L}_{\text{adv}}(G, D) &= D(G(z)) - D(x) \\ G^*_{s\to t} &= \mathbb{E}_{z\sim p_z(z), x\sim \mathcal{D}_t} \arg\min_G \max_D \mathcal{L}_{\text{adv}}(G, D). \end{aligned} \tag{1}$$

Previous work [32] shows that this works well when the target dataset size exceeds 1000 training samples. However, in the extreme few-shot setting, this method overfits, as the discriminator can memorize the few examples and force the generator to reproduce them. This is shown in Fig. 2, where we see collapse after tuning the source model (top row) to the few-shot target dataset (middle row).

To prevent overfitting to generate diverse and realistic images (Fig. 2, bottom row), we propose a new cross-domain consistency loss (Sec. 3.1), which actively uses the original source generator to regularize the tuning process, and a "relaxed" discriminator (Sec. 3.2), which encourages different levels of realism over different regions in the latent space. Our approach is shown in Fig. 3.

3.1. Cross-domain distance consistency

A consequence of overfitting during adaptation is that relative distances in the source domain are not preserved. As seen in Fig. 2, the visual appearance between z_1 and z_2 collapses, disproportionately relative to z_1 and z_3, which remain perceptually distinct. We hypothesize that enforcing preservation of relative pairwise distances, before and after adaptation, will help prevent collapse.

To this end, we sample a batch of $N + 1$ noise vectors $\{z_n\}_0^N$, and use their pairwise similarities in feature space to construct N-way probability distributions for each image. This is illustrated in Fig. 3 from the viewpoint of z_0.

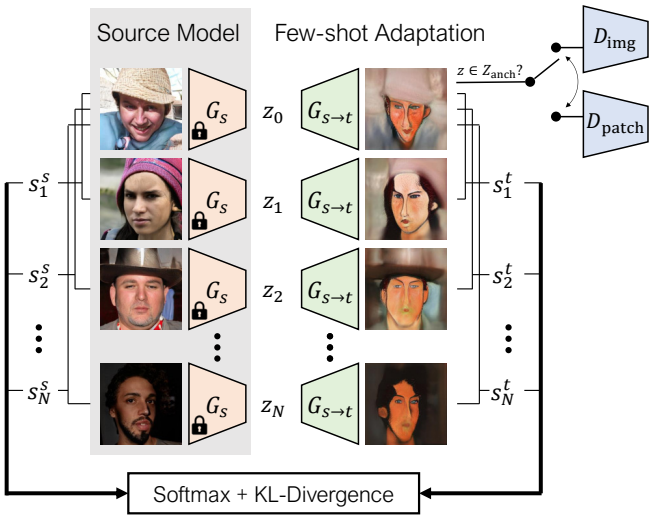

Figure 3: Our approach contains two key elements. (1) Cross-domain consistency loss $\mathcal{L}_{\text{dist}}$ aims to preserve the relative pairwise distances between source and target generations. In this case, the relative similarities between synthesized images from z_0 and other latent codes are encouraged to be similar. (2) Relaxed realism is implemented by using two discriminators, D_{img} for noise sampled from the anchor region (z_{anch}) and D_{patch} otherwise.

The probability distribution for the i^{th} noise vector, for the source and adapted generators is given by,

$$\begin{aligned} y_i^{s,l} &= \text{Softmax}\big(\{\text{sim}(G_s^l(z_i), G_s^l(z_j))\}_{\forall i\neq j}\big) \\ y_i^{s\to t,l} &= \text{Softmax}\big(\{\text{sim}(G_{s\to t}^l(z_i), G_{s\to t}^l(z_j))\}_{\forall i\neq j}\big), \end{aligned} \tag{2}$$

where sim denotes the cosine similarity between generator activations at the l^{th} layer. We are inspired by recent methods in contrastive learning [22, 4, 7], which converts similarities into probability distributions for unsupervised representation learning, as well as perceptual feature losses [10, 5, 26], which show that activations on multiple layers on discriminative networks help preserve similarity. We encourage the adapted model to have similar distributions to the source, across layers and image instances by using KL-divergence:

$$\mathcal{L}_{\text{dist}}(G_{s\to t}, G_s) = \mathbb{E}_{\{z_i\sim p_z(z)\}} \sum_{l,i} D_{KL}\big(y_i^{s\to t,l} || y_i^{s,l}\big). \tag{3}$$

While correspondence helps prevent collapse, we also modify the adversarial loss to further prevent overfitting.

3.2. Relaxed realism with few examples

With a very small target data size, the definition of what constitutes a "realistic" sample becomes increasingly overconstrained, as the discriminator can simply memorize the few-shot target training set. We note that the few training

images only form a small subset of the desired distribution and extend this notion to the latent space. We define "anchor" regions, $\mathcal{Z}_{\text{anch}} \subset \mathcal{Z}$, which form a subset of the entire latent space. When sampled from these regions, we use a full image discriminator D_{img}. Outside of them, we enforce adversarial loss using a patch-level discriminator D_{patch},

$$\mathcal{L}'_{\text{adv}}(G, D_{\text{img}}, D_{\text{patch}}) = \mathbb{E}_{x\sim\mathcal{D}_t}\big[\mathbb{E}_{z\sim Z_{\text{anch}}}\mathcal{L}_{\text{adv}}(G, D_{\text{img}}) + \mathbb{E}_{z\sim p_z(z)}\mathcal{L}_{\text{adv}}(G, D_{\text{patch}})\big]. \tag{4}$$

To define the anchor space, we select k random points, corresponding to the number of training images, and save them. We sample from these fixed points, with a small added Gaussian noise ($\sigma = .05$). We use shared weights between the two discriminators by defining D_{patch} as a subset of the larger D_{img} network [9, 38]; using internal activations correspond to patches on the input. The size depends on the network architecture and layer. We read off a set of layers, with effective patch size ranging from 22×22 to 61×61.

3.3. Final Objective

Our final objective consists of just these two terms: $\mathcal{L}'_{\text{adv}}$ for the appearance of the target and $\mathcal{L}_{\text{dist}}$, which directly leverages the source model to preserve structural diversity:

$$G^*_{s\to t} = \arg\min_G \max_{D_{\text{img}}, D_{\text{patch}}} \mathcal{L}'_{\text{adv}}(G, D_{\text{img}}, D_{\text{patch}}) + \lambda\,\mathcal{L}_{\text{dist}}(G, G_s). \tag{5}$$

The patch discriminator gives the generator some additional freedom on the structure of the image. The adapted generator is directly incentivized to borrow the domain structure from the source generator, due to the cross-domain consistency loss. As shown in the top and bottom rows in Fig. 2, the model indeed discovers *cross-domain correspondences* between source and target domains.

We use the StyleGANv2 architecture[1] [13], pre-trained on a large dataset (e.g. FFHQ [12]) as our source model. We use a batch size of 4. Empirically, we find that a high λ, from 10^3 to 10^4, to work well. Additional training details can be found in the supplementary.

4. Experiments

We explore different `source` $\rightarrow$ `target` adaptation settings to analyze the effectiveness of our approach in preserving part-level correspondences between images generated from G_s and $G_{s\to t}$. We also investigate what kinds of correspondences emerge when the source and target domains are unrelated.

Baselines: We compare to baselines, which similar to ours, adapt a pre-trained source model to a target domain with limited data. (i) Transferring GANs (TGAN) [32]: fine-tunes a pre-trained source model to a target domain with the same objective used to train the source model; (ii) Batch Statistics Adaptation (BSA) [21]: only adapts the scale and shift parameters of the model's intermediate layers; (iii) MineGAN [30]: for a given pre-trained source (e.g. MNIST 0-8) and target (e.g. 9) domain, it transforms the original latent space of source to a space more relevant for the target (e.g. mapping all 0-8 regions to 4, being more similar to 9); (iv) Freeze-D [19]: freezes the high resolution discriminator layers during adaptation; (v) Non-leaking data augmentations [11, 37]: uses adaptive data augmentations (TGAN + ADA) in a way that does not leak into the generated results; (vi) EWC [15]: extends the idea of Elastic Weight Consolidation [14] for adapting a source model to a target domain, by penalizing large changes to *important* weights (estimated via Fisher information) in the source model.

Datasets: We use source models trained on five different datasets: (i) Flickr-Faces-HQ (FFHQ) [12], (ii) LSUN Church, (iii) LSUN Cars, and (iv) LSUN Horses [35]. We explore adaptation to the following target domains: (i) face caricatures, (ii) face sketches [29], (iii) face paintings by Amedeo Modigliani [34], (iv) FFHQ-babies, (v) FFHQ-sunglasses, (vi) landscape drawings, (vii) haunted houses, (viii) Van Gogh's house paintings, (ix) wrecked/abandoned cars. We operate on 256×256 resolution images for both the source and target domains. Adaptation is done on 10 images from the target domain, unless stated otherwise.

4.1. Quality and Diversity Evaluation

Qualitative comparison Fig. 4 shows results on two target domains using different methods, all of which start from the same source model pre-trained on FFHQ (Source). We observe that TGAN strongly overfits to the available training data, as was the case for Amedeo paintings (Fig. 2). Using adaptive data augmentations (TGAN + ADA) has little to no effect on sketches, and further degrades the quality for caricatures, where augmentations (e.g. 90° rotations) leak into the generated images. FreezeD, MineGAN and EWC perform better than TGAN by generating slightly more diverse images. However, the diversity is only introduced through minor modifications among the few captured modes. For example, (i) the caricature results for EWC show multiple generations capturing the same person with different expression/hairlines; (ii) the results on sketches using MineGAN depict a person with similar attributes in multiple generations. Our method better captures the distribution of caricatures and sketches, and generates diverse and realistic images containing objects which do not appear in the training images (e.g. hats, in sketches). This is because our method is flexible enough to not penalize the generated images which do not adhere to the 10 training samples.

[1] https://github.com/rosinality/stylegan2-pytorch

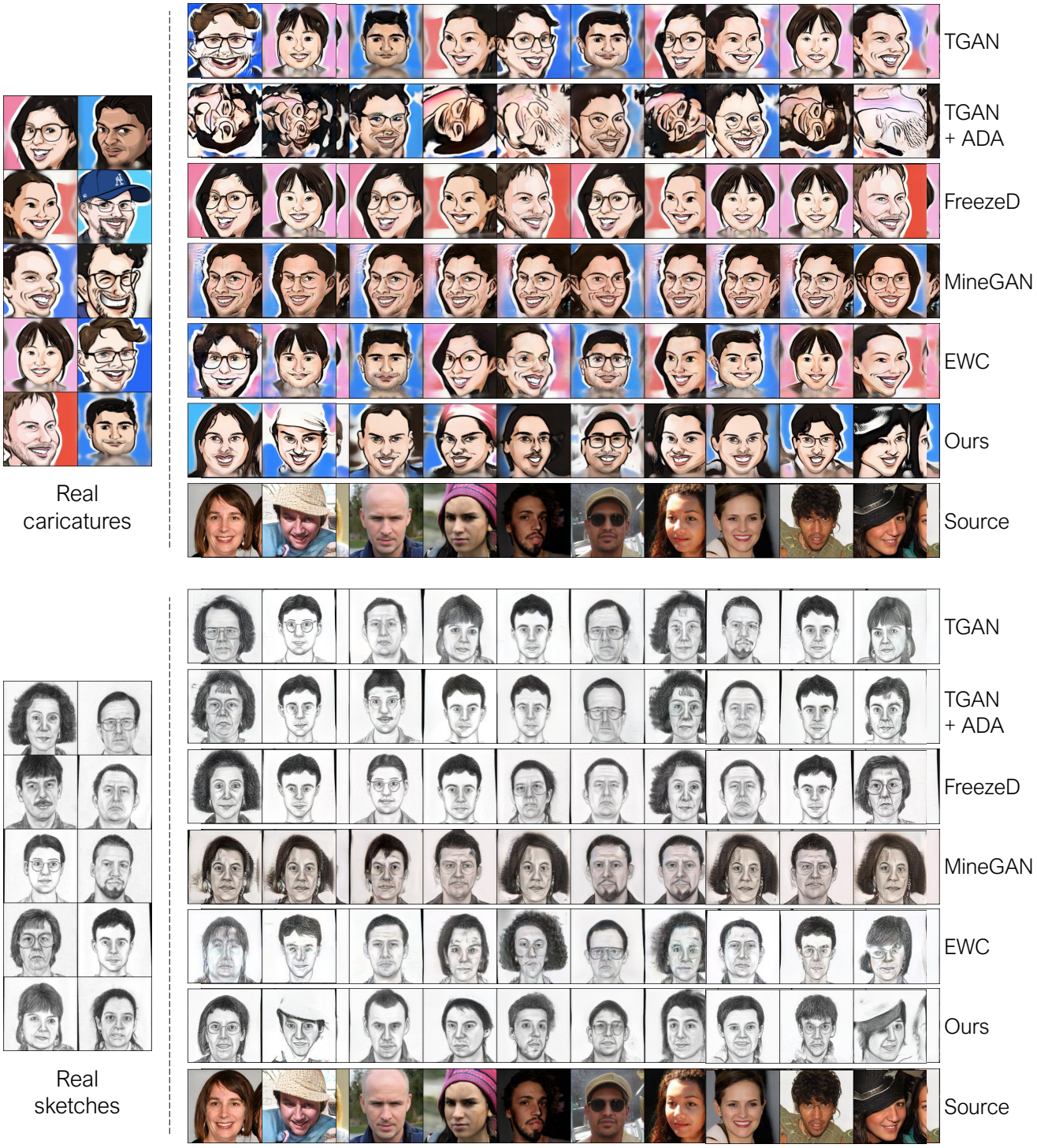

Figure 4: 10-shot image synthesis results for different methods, which start from the same source model (bottom). Keeping the noise vectors same (across columns), we observe that the baselines either overfit, or only capture a few modes in the target domain. Our method generates higher quality and more diverse results which better correspond to the source domain images generated from the same noise.

Quantitative comparison The original Sketches, FFHQ-babies, and FFHQ-sunglasses datasets roughly contain 300, 2500, and 2700 images, respectively. To simulate a few-shot setting, we randomly sampled 10 images from each

Figure 5: Effect of different components of our method. Absence of $\mathcal{L}_{dist}$ makes some properties of the images (e.g. hairline) look very similar. Application of only one of D_{img} and D_{patch} degrades the image quality by distorting the face structure.

Figure 6: Visualizing the emerging correspondence in different adaptation settings. The generations from the domain of Van Gogh houses resemble the building structure from the source Church images. The generated wrecked/abandoned cars preserve the source car's body and pose. Generations from FFHQ → Sunglasses learn to add sunglasses to people's faces.

dataset to train our model. For evaluation purposes, however, we can use the entire dataset to measure how well our generated images model the true distribution.

Table 1 shows the FID [8] scores. Our method significantly outperforms all baselines for Babies and Sunglasses. For domains with limited data, however, the FID score would not reflect the overfitting problem.

Ideally, we wish to assess the number of visually distinct images an algorithm can generate. In the worst case, the algorithm will simply overfit to the original k training images. To capture this, we first generate 1000 images and assign them to one of the k training images, by using lowest LPIPS distance [36]. We then compute the average pairwise LPIPS distance within members of the same cluster and then average over the k clusters. A method that reproduces the original images exactly will have a score of zero by this metric. Table 2 summarizes the distances for different baselines over three target domains. We see that

	Babies	Sunglasses	Sketches
TGAN [32]	104.79 ± 0.03	55.61 ± 0.04	53.41 ± 0.02
TGAN+ADA [11]	102.58 ± 0.12	53.64 ± 0.08	66.99 ± 0.01
BSA [21]	140.34 ± 0.01	76.12 ± 0.01	69.32 ± 0.02
FreezeD [19]	110.92 ± 0.02	51.29 ± 0.05	46.54 ± 0.01
MineGAN [30]	98.23 ± 0.03	68.91 ± 0.03	64.34 ± 0.02
EWC [15]	87.41 ± 0.02	59.73 ± 0.04	71.25 ± 0.01
Ours	**74.39 ± 0.03**	**42.13 ± 0.04**	**45.67 ± 0.02**

Table 1: FID scores (↓) for domains with abundant data. Standard deviations are computed across 5 runs.

our method consistently achieves higher average LPIPS distances, indicating more distinct images being generated. We also visualize the cluster centers and their members, to see if they are semantically meaningful. See supp. for details.

What role do different components of our method play? We use caricature as the target domain, and first study the effect of our framework with and without $\mathcal{L}_{\text{dist}}$; see Fig. 5.

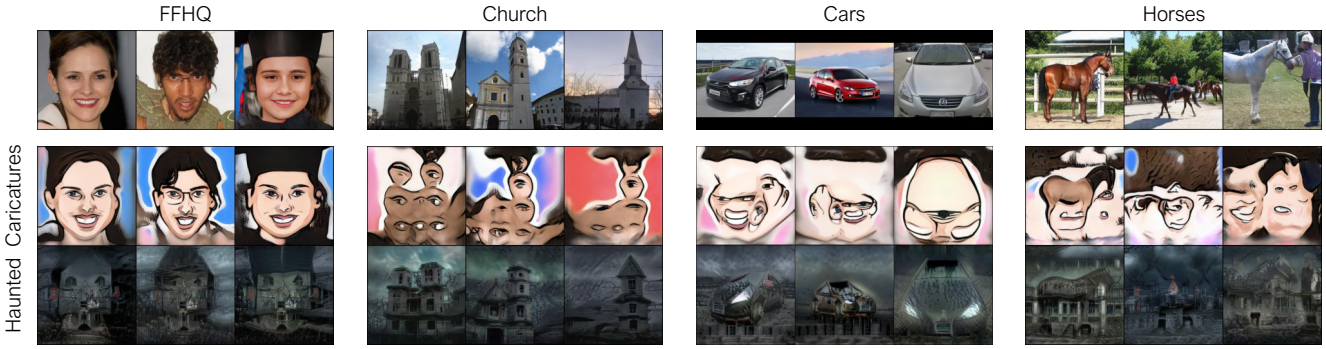

Figure 7: Visualizing the effect of unrelated source domains. In most cases, accurate modeling of a target distribution is not feasible when starting from an unrelated source domain. However, our method still discovers a correspondence on a part-level between the two domains, where different parts of the source (car's tires) correspond to parts of the target (caricature's eyes).

	Caricatures	Amedeo's paintings	Sketches
TGAN [32]	0.39 ± 0.06	0.41 ± 0.03	0.39 ± 0.03
TGAN+ADA[11]	0.50 ± 0.05	0.51 ± 0.04	0.41 ± 0.05
BSA [21]	0.35 ± 0.01	0.39 ± 0.04	0.35 ± 0.01
FreezeD [19]	0.37 ± 0.01	0.40 ± 0.03	0.39 ± 0.03
MineGAN [30]	0.39 ± 0.07	0.42 ± 0.03	0.40 ± 0.05
EWC [15]	0.47 ± 0.03	0.52 ± 0.03	0.42 ± 0.03
Ours	$\mathbf{0.53 \pm 0.01}$	$\mathbf{0.60 \pm 0.01}$	$\mathbf{0.45 \pm 0.02}$

Table 2: Intra-cluster pairwise LPIPS distance (↑). Standard deviation is computed across the k clusters (k = no. of training samples).

We see that leaving out $\mathcal{L}_{\text{dist}}$ reduces diversity among the generations, all of which have very similar head structure and hair style. We next study the different ways we enforce realism. What happens if we keep $\mathcal{L}_{\text{dist}}$, but use image-level adversarial loss through D_{img} on *all* generations? 'Ours w/ only D_{img}' results reveal the problem of mode collapse at the part level (same blue hat appears in multiple generations) and the phenomenon where some results are only slight modifications of the same mode (same girl with and without the blue hat). Could we then only use D_{patch} to enforce patch-level realism on *all* generations? 'Ours w/ only D_{patch}' shows the results, where we observe more diversity, but poorer quality compared to 'Ours w/ only D_{img}'. This is because the discriminator never gets to see a *complete* caricature image, and consequently does not learn the part-level relations which makes a caricature look realistic. 'Ours' combines all the ideas, resulting in generations which are diverse and realistic at both the part and image level.

4.2. Analyzing source ↔ target correspondence

As mentioned before, our method captures, to a large extent, the correspondence between the images from source and target domains, and preserves it during the adaptation process. In this section, we study this property in detail in different source to target adaptation settings.

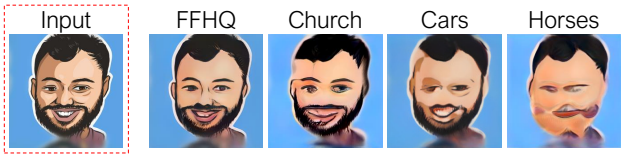

Figure 8: Embedding an unseen caricature image (left), and visualizing its reconstruction from models adapted from different source to the caricature domain.

Related source/target domains Our method builds around the idea of discovering correspondence between a source and a target domain. But how well does that property actually hold? We first consider the FFHQ → Caricatures/sketches setting in Fig. 4,and analyze the baselines' results. As seen previously in Sec. 4.1, the generated images mostly overfit to the training samples and are unable to borrow anything other than rough pose from the corresponding source. Using our method, on the other hand, we observe that the content of the natural source faces, as well as any of the accessories (e.g. hats/sunglasses), are preserved in both the sketch and caricature domain, depicting a much cleaner correspondence. We further study this for other source→ target settings: (i) Church → Van Gogh houses, (ii) Cars → Abandoned cars, (iii) FFHQ → Sunglasses, as shown in Fig 6. When the source and target domains have similar semantics, the results generated from the same noise vectors in respective domains have clear correspondence.

Unrelated source/target domains We adapt four source models (FFHQ, Church, Cars, and Horses) to two target domains (Caricatures and Haunted houses) and present the results in Fig. 7. For FFHQ → Caricatures and Church → Haunted, the generated results from the adapted model mimic the target domain appearance and reflect correspondences with the source. For all the remaining scenarios, the adaptations do not capture the target distribution accurately. However, some part-level correspondences still

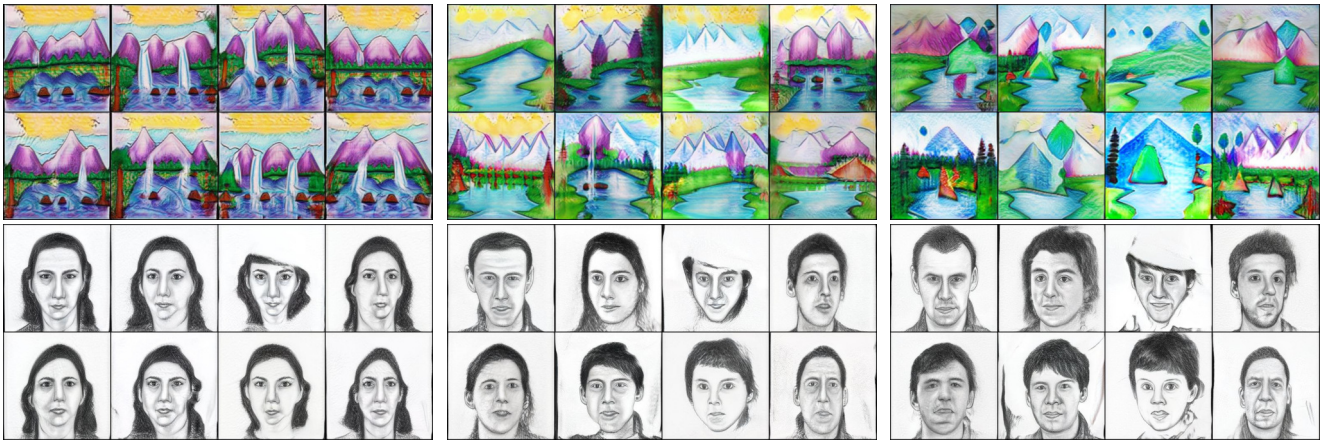

Figure 9: **Effect of training data size:** Even with just a single image, our method can capture it in different modes through the generations. Increasing the number of training samples results in more diversity in the generated sketches and landscapes.

emerge in these settings. E.g., (i) Church → Caricatures: windows/doors of the Church roughly map to eyes of the caricature; (ii) Cars → Caricatures: wheels/bumpers of the cars adapt to represent eyes/mouth of caricatures respectively; (iii) Cars → Haunted: moon in the haunted houses (see supp. for the 10 images used) maps to the headlights, lighting them up; (iv) Horses → Haunted: bottom of the horse legs adapt to doors of haunted houses.

Quantitative analysis of source/target relevance We translate four source models to four target domains: caricatures, haunted houses, landscape drawings, and abandoned cars. We then test how well a translated model can embed an *unseen* image from the respective target domain, E.g., after translating the four source models to the caricature domain, we use a new caricature and embed it into the four models using Image2StyleGAN [1]. The results are shown in Fig. 8. In Table 3, for each domain, we report the average similarity scores between five unseen inputs and their reconstructions. We see that FFHQ, Church, and Cars best reconstruct images from caricatures, haunted houses, and abandoned cars respectively which aligns with our intuition.

	Caricature	Haunted house	Wrecked car
FFHQ	**0.158 ± 0.045**	0.645 ± 0.024	0.643 ± 0.012
Church	0.294 ± 0.077	**0.599 ± 0.028**	0.621 ± 0.032
Cars	0.233 ± 0.106	0.635 ± 0.031	**0.606 ± 0.057**
Horses	0.299 ± 0.083	0.631 ± 0.028	0.619 ± 0.038

Table 3: Relevance of source and target domains, measured via LPIPS (↓) between an unseen image and its reconstruction. Better the reconstruction → more similar domains.

4.3. Effect of target dataset size

So far, all results are generated with 10 training images per target domain. We now explore how the dataset size affects the quality and diversity of the generated images. We consider two adaptation setups, Church → Landscape drawings and FFHQ → Sketches, and present results in 1-shot, 5-shot and 10-shot settings; see Fig. 9. The real images used in these settings can be found in the supplementary.

In 1-shot, our method introduces small variations to the single target image, for example the lady appears in different poses in the generated sketches, and the mountains and waterfall have different structures. The diversity of the results increases with 5 training images. The sketches now reflect distinct identities. Further increasing the number of training samples (10-shot) introduces more details for the sketch domain, and generates more diverse landscapes.

5. Conclusion and Limitations

We proposed to adapt a pretrained GAN learned on a large source domain to a small target domain by discovering cross-domain correspondences. While our method generates compelling results, it is not without limitations. Cars → Abandoned cars in Fig. 6 depicts an example, where the color of the red car changes to orange in its abandoned form, likely because of the existence of an orange car (and no red one) in the 10 training images. FFHQ → Sunglasses depicts another example, where a blonde hair turns dark with sunglasses. These show that there is a need for discovering better correspondence between the source and target domains, which will lead to more diverse generations. Nevertheless, we believe this work takes an important step towards creating more data-efficient generative models, demonstrating that existing source models can be leveraged in an effective way to model new distributions with less data.

Acknowledgements: We thank Daichi Ito for the beautiful caricatures. Part of the work was supported through NSF CAREER IIS-1751206 and Adobe Data Science Research Award.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Int. Conf. Comput. Vis.*, 2019.

[2] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *Adv. Neural Inform. Process. Syst.*, 2017.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[5] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4829–4837, 2016.

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[7] 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.

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Adv. Neural Inform. Process. Syst.*, 2017.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[11] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Adv. Neural Inform. Process. Syst.*, 2020.

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[13] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*, 2019.

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

[15] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In *Advances in Neural Information Processing Systems*, 2020.

[16] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Int. Conf. Comput. Vis.*, 2019.

[17] Shaohui Liu, Xiao Zhang, Jianqiao Wangni, and Jianbo Shi. Normalized diversification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[18] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[19] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze discriminator: A simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020.

[20] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[21] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Int. Conf. Comput. Vis.*, 2019.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[23] Kuniaki Saito, Kate Saenko, and Ming-Yu Liu. Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. *arXiv preprint arXiv:2007.07431*, 2020.

[24] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Adv. Neural Inform. Process. Syst.*, 2017.

[25] Ngoc-Trung Tran, Tuan-Anh Bui, and Ngai-Man Cheung. Dist-gan: An improved gan using distance constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[26] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, volume 1, page 4, 2016.

[27] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[28] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Adv. Neural Inform. Process. Syst.*, 2016.

[29] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(11):1055–1967, 2009.

[30] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[31] Yaxing Wang, Salman Khan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan. Semi-supervised learning for few-shot image-to-image translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[32] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Eur. Conf. Comput. Vis.*, 2018.

[33] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. *arXiv preprint arXiv:1901.09024*, 2019.

[34] Jordan Yaniv, Yael Newman, and Ariel Shamir. The face of art: landmark detection and geometric style in portraits. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019.

[35] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, 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.

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[37] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *arXiv preprint arXiv:2006.10738*, 2020.

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis.*, 2017.

[39] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Adv. Neural Inform. Process. Syst.*, 2017.