Improving Diversity in Zero-Shot GAN Adaptation with Semantic Variations

Seogkyu Jeon[1]* Bei Liu[2] Pilhyeon Lee[1] Kibeom Hong[1,3] Jianlong Fu[2] Hyeran Byun[1]†
[1]Yonsei University [2]Microsoft Research Asia [3]SwatchOn
jone9312@yonsei.ac.kr

Abstract

Training deep generative models usually requires a large amount of data. To alleviate the data collection cost, the task of zero-shot GAN adaptation aims to reuse well-trained generators to synthesize images of an unseen target domain without any further training samples. Due to the data absence, the textual description of the target domain and the vision-language models, e.g., CLIP, are utilized to effectively guide the generator. However, with only a single representative text feature instead of real images, the synthesized images gradually lose diversity as the model is optimized, which is also known as mode collapse. To tackle the problem, we propose a novel method to find semantic variations of the target text in the CLIP space. Specifically, we explore diverse semantic variations based on the informative text feature of the target domain while regularizing the uncontrolled deviation of the semantic information. With the obtained variations, we design a novel directional moment loss that matches the first and second moments of image and text direction distributions. Moreover, we introduce elastic weight consolidation and a relation consistency loss to effectively preserve valuable content information from the source domain, e.g., appearances. Through extensive experiments, we demonstrate the efficacy of the proposed methods in ensuring sample diversity in various scenarios of zero-shot GAN adaptation. We also conduct ablation studies to validate the effect of each proposed component. Notably, our model achieves a new state-of-the-art on zero-shot GAN adaptation in terms of both diversity and quality.

1. Introduction

In recent years, deep generative models, especially generative adversarial networks (GANs) [9], have shown dramatic advancements by successfully mimicking the real distribution of images [3, 13, 6]. However, as diagnosed in the literature [47, 25], building powerful generative models requires a huge number of visual samples as well as expensive training costs. This essentially restricts the applicability of the models to domains where it is prohibitively expensive or even infeasible to collect sufficient data, such as medical images or artworks of specific artists. To alleviate the limitation, researchers give their attention to the task of GAN adaptation, which aims to reuse the representation power of well-trained generators for synthesizing images of a target domain. To this end, existing works manage to transfer the generation capability of pre-trained GANs to unseen target domains by exploiting a tiny dataset [40] with only a few visual samples (*i.e.*, few-shot) [39, 22, 23, 28, 34, 29], or even no data at all (*i.e.*, zero-shot) [8]. This paper focuses on the zero-shot setting, where a generative model pre-trained on a source dataset is supposed to be adapted to an unseen target domain that contains no visual samples for training.

In order to perform adaptation with no accessibility to data of the target domain, the previous work hinges on the powerful vision-language model, *i.e.*, CLIP [32], which learns the shared latent space between vision and text modalities. Specifically, StyleGAN-NADA [8] embeds two textual prompts respectively describing the source and target domains into the CLIP space and derives the difference vector between them. Considering the difference vector to be the guiding direction, the generated images gradually step toward the target domain. Eventually, the generator is able to synthesize visually plausible images of the target domain even without seeing any samples of the domain.

However, the adapted model with only a single guiding direction suffers from *mode collapse*. That is, the generated target samples share the same characteristics without distinction. For instance, the generated faces under the "Photo-to-Pixar" scenario have exactly the same attributes, *e.g.*, emotional expression, slightly opened mouth, dark hair (figure 1 (a) center). In another example, the results under the "Dog-to-Cat" scenario exhibit nearly identical cat faces with few differences (figure 1 (b) center). These problems come from the one-to-one mechanism of the CLIP text encoder; given a single guiding direction, the adapted generator is unable to handle the diversity of the target domain. Intuitively, the target textual prompt provides the most general

[*] Work done during his internship at Microsoft Research Asia
[†] Corresponding author

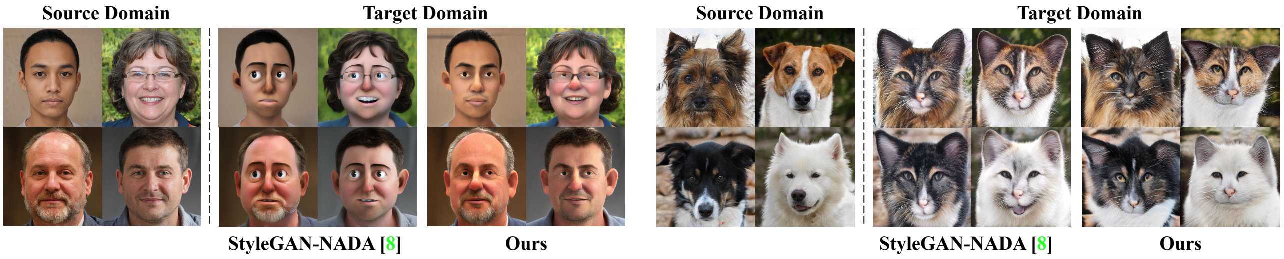


Figure 1. Illustration of our motivation. For two adaptation scenarios, *i.e.*, "Photo-to-Pixar" and "Dog-to-Cat", we present source domain images and the corresponding generated images of the target domain by StyleGAN-NADA [8] and ours.

information about the target domain while there are an infinitely large number of semantic variations underneath. For instance, the target domain "Cat" implicitly covers a variety of species with diverse characteristics such as a smiling Sphynx and a brown Scottish Fold. Hence, relying solely on a single target description fails to exploit the inherent semantic variations, and it is crucial to model a one-to-many relation for enhancing sample diversity after adaptation.

In this paper, we explore semantic variations of the given text prompt of the target domain with a novel two-stage framework in order to alleviate the mode collapse problem. In the first stage, to discover the variations, we impose a set of learnable perturbations on the target text embedding in the CLIP space. They are encouraged to be orthogonal to each other for redundancy reduction yet not to disturb the original semantics of the target domain. The obtained variations are then used to compute guiding directions. In the second stage, to improve sample diversity using multiple guidances, we introduce a novel *directional moment loss*. It effectively aligns image-updating directions with the guidances by matching their first and second moments.

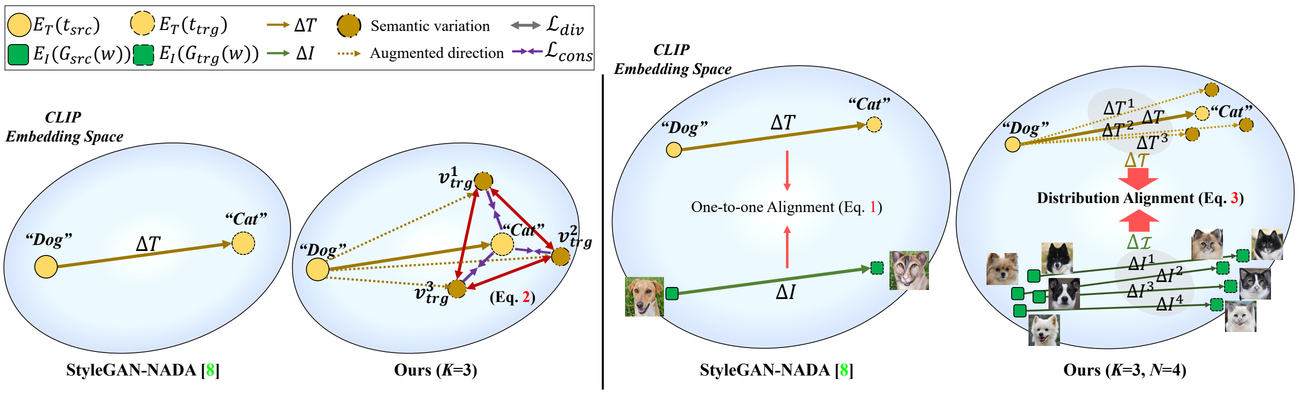
In addition, we propose a *relation consistency loss* to better sustain the knowledge of the generator learned from the source domain. Ideally, the relation between two generated images should remain the same during adaptation to ensure consistency of semantic information. From this motivation, the relation consistency loss is designed to minimize the distribution gap between the inter-image relations of the source and target domains. Further, we employ the elastic weight consolidation [19] to suppress excessive changes of important parameters of the generator during adaptation. This prevents our model from losing the strong content representability of the generator during the adaptation, thereby preserving the original content information.

Equipped with the proposed components, our model is able to generate images with diverse semantic variations of the target domain, while successfully preserving the original semantic information of the source domain. The superiority of our method over the previous work is clearly showcased in Figure 1. Through extensive experiments on various adaptation scenarios, we demonstrate the effectiveness of each component of our model. Moreover, our model achieves a new state-of-the-art on zero-shot GAN adaptation in terms of quality as well as diversity.

2. Related Works

Few-shot GAN adaptation. In the last decade, research on deep generative models has achieved remarkable advances and they are now capable of almost completely mimicking the distributions of real images [3, 13, 14, 12]. However, on the dark side, they require an excessively large amount of real images for effective and stable training from scratch. Constructing a large-scale well-refined training dataset is excessively costly and laborious, and even unavailable in some domains, *e.g.*, artworks. To this end, several studies [11, 47, 38, 37, 44, 49] are proposed to accomplish data-efficient training with a small number of training samples provided (*e.g.*, 10^3 to 10^4).

Despite their achievements, the studies still struggle in a more restrictive setting where only a few samples less than 10 are accessible. Due to the formidable data scarcity, the generator is prone to overfitting, *i.e.*, memorizing only some training samples, thereby losing diversity and falling into mode collapse. To tackle the problem, the task of few-shot GAN adaptation [26, 39, 22, 23, 28, 34, 40, 29, 46, 41, 48, 50, 17] arises to adapt well-trained generative models to the target domain. As the knowledge of the target domain is largely limited, pre-trained GANs are generally leveraged to distill the diverse content information learned from the large-scale dataset. MineGAN [39] introduces the mining network to identify beneficial knowledge for the target domain generation. Meanwhile, Li et al. [22] preserve the weights of the generative models with elastic weights consolidation [19] based on Fisher information. Ojha et al. [29] propose a GAN adaptation framework with a cross-domain correspondence loss and a relaxed discriminator. RSSA [41] proposes a relaxed spatial consistency method that encourages the generator maintain the self-correlation and the inter-sample spatial correlation. DCL [48] proposes a contrastive learning framework to enhance visual quality and ameliorate diversity degradation.



(a) Semantic Variation Learning

(b) Directional Moment Loss

Figure 2. Illustration of the proposed methods. (left) semantic variation learning in CLIP space (right) the directional moment loss. Here K denotes the number of augmented semantic variations of the target text v^i_{trg}, while N is the batch size of generated samples.

Zero-shot GAN adaptation. Recently, CLIP [32] has brought a significant impact on the computer vision fields, showing impressive performance as well as robustness in the zero-shot classification task. Thanks to its powerful cross-modal representation, recent research [31, 7, 4, 27, 43, 8, 31, 20, 16] have actively attempted to exploit the pre-trained CLIP for generative tasks.

Among others, bringing the power of CLIP to GAN, StyleGAN-NADA [8] enables GAN adaptation only with the textual descriptions of the target domain without any training images, *i.e.*, zero-shot GAN adaptation. Using the guiding direction obtained from the textual domain descriptions, it adapts the generator so that the generated images move in accordance with the guidance in the CLIP space. As a result, it can synthesize samples of the target domain by relying on only textual descriptions, not image samples. Unlike the previous latent manipulation method [31] where available modifications are constrained in the domain of the pre-trained generator, zero-shot GAN adaptation can perform out-of-domain manipulation by directly optimizing the generator parameters.

However, as shown in Figure 1, StyleGAN-NADA fails to capture diverse semantic variations of the target text, resulting in the mode collapse of generated samples. We conjecture that this is due to its adaptation process that relies heavily on a single target domain description. To handle this problem, we propose to discover semantic variations of the target text in CLIP space, which enables generating diverse samples of the target domain in the zero-shot setting.

3. Proposed Methods

3.1. Baseline

Our baseline for text-driven GAN adaptation is similar to StyleGAN-NADA [8] except that we do not use complex layer selection. Its architecture basically follows StyleGAN2 [14] which consists of a mapping network and a generator G_{src}. The mapping network is trained to embed a latent code from the prior distribution into the disentangled latent space $\mathcal{W}$. The generator G_{src} takes the converted latent code $w \in \mathbb{R}^{B\times D_w}$ as input to generate RGB images $G_{src}(w) \in \mathbb{R}^{N\times 3\times H\times W}$ of the training domain. Here D_w is the dimension of the latent space, N denotes the batch size, and (H, W) indicates the size of generated images.

The main training objective is to adapt the pre-trained generator G_{src} on a source domain (*e.g.*, cat) to synthesize the images of a target domain (*e.g.*, dog) using the descriptions of the source and target domains, *i.e.*, t_{src} and t_{trg}, as the text prompt. Note that G_{src} is sufficiently optimized to generate realistic samples of the source domain. The target domain generator G_{trg} is initialized by the parameters of G_{src} and optimized during training, whereas G_{src} and the mapping network remain frozen. By doing so, we can generate source samples $G_{src}(w)$ and target samples $G_{trg}(w)$ from the same latent code w to estimate the image-level relation between two domains. To convey the learned knowledge of G_{src} to G_{trg}, the directional loss $\mathcal{L}_{dir}$ is designed to align the direction between source and target images with the text direction in the CLIP embedding space by maximizing their cosine similarities as follows.

$$\mathcal{L}_{dir} = \frac{1}{N}\sum_{n=1}^{N}\left[1 - \frac{\Delta I^n \cdot \Delta T}{\|\Delta I^n\|\,\|\Delta T\|}\right], \tag{1}$$
$$\text{where } \Delta T = E_T(t_{trg}) - E_T(t_{src})$$
$$\text{and } \Delta I^n = E_I(G_{trg}(w^n)) - E_I(G_{src}(w^n)).$$

Here N is the mini-batch size, while E_T and E_I respectively denote the text encoder and the image encoder of the pre-trained CLIP [32] that share the same embedding space with its dimension of D. It is worth noting that the directional loss encourages *every* image sample to be updated in the same direction with the text guidance, *i.e.*, one-to-one alignment (see Figure 2 (b) left).

3.2. Motivation

As the CLIP text encoder is in nature deterministic, the directional loss $\mathcal{L}_{dir}$ is computed with only a single direction toward the representative feature of the target domain, *i.e.*, ΔT. Consequently, all generated image samples are updated in the same direction and encouraged to share typical characteristics, while gradually diminishing the diversity during the adaptation. Utilizing some textual templates to decorate the target text can be deemed an intuitive solution, but manually defining templates for each specific target domain is heuristic and less generalizable. Instead, we propose to augment the features by exploring *semantic variations*, in order to alleviate the mode collapse. Here, the semantic variations denote feature vectors that are semantically consistent with the target text while being capable of expressing diverse characteristics.

The overall pipeline of our method consists of two stages. In the first stage, we search for diverse semantic variations to augment the target text feature $E_T(t_{trg})$ while preserving its original information. In the second stage, we guide the target generator G_{trg} with the augmented text directions while maintaining sample diversity.

3.3. Semantic Variation Learning

To find semantic variations in the CLIP space, we first prepare K learnable vectors $\{z^i\}_{i=1}^{K}$, where K denotes the number of variations and the vectors share the same space with the target text feature, *i.e.*, $z^i \in \mathbb{R}^D$. Thereafter, each vector serves as an additive perturbation on the target text feature and learns a useful semantic variation that can enhance the diversity but does not disturb the original semantics of the target domain text. Concretely, the vectors are optimized with two loss functions as follows.

$$\begin{aligned}\mathcal{L}_{cons} &= \frac{1}{K}\sum_{i=1}^{K}\Big[1 - \frac{E_T(t_{trg})\cdot v^i_{trg}}{\|E_T(t_{trg})\|\,\|v^i_{trg}\|}\Big],\\ \mathcal{L}_{div} &= \binom{K}{2}^{-1}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\left|\frac{z^i\cdot z^j}{\|z^i\|\,\|z^j\|}\right|,\end{aligned} \tag{2}$$

where $v^i_{trg} = E_T(t_{trg}) + \epsilon\frac{z^i}{\|z^i\|}$ denotes the perturbed text feature, *i.e.*, the semantic variation, and ϵ is a hyperparameter determining the perturbation strength. $\mathcal{L}_{cons}$ is a semantic consistency loss that prevents unintended deviation of the original semantic by regularizing the cosine distance between the original text feature $E_T(t_{trg})$ and its semantic variations v^i_{trg}. On the other hand, $\mathcal{L}_{div}$ is a semantic diversity loss that prevents the perturbation vectors from learning redundant information by encouraging orthogonality for all combinations of z^*.

To summarize, we search the semantic variations $\{v^i_{trg}\}_{i=1}^{K}$ by optimizing $\{z^i\}_{i=1}^{K}$ with the weighted sum of the losses $\mathcal{L}_{S1} = \mathcal{L}_{cons} + \lambda_{div}\mathcal{L}_{div}$, where λ_{div} is a weighting factor. Our semantic variation learning is depicted in Figure 2 (a).

3.4. Directional Moment Loss

After searching the semantic variations $\{v^i_{trg}\}_{i=1}^{K}$, we utilize them as a kind of augmentation to guide G_{trg} with multiple directions between the source and target texts. To encourage G_{trg} to learn the diversity from a single target text and its semantic variations, we propose a novel *directional moment loss*. In specific, we first compute the text direction from the source text feature $E_T(t_{src})$ to each semantic variation v^i_{trg} as $\Delta T^i = v^i_{trg} - E_T(t_{src})$. Then we compose the text direction set with the original direction ΔT and perturbed ones ΔT^* as $\Delta\mathcal{T} = [\Delta T; \Delta T^1; \ldots; \Delta T^K]^\top \in \mathbb{R}^{(K+1)\times D}$, where D denotes the channel dimension. Meanwhile, the image direction set can be obtained by composing the image directions within the batch: $\Delta\mathcal{I} = [\Delta I^1; \Delta I^2; \ldots; \Delta I^N]^\top \in \mathbb{R}^{N\times D}$. We design a directional moment loss to minimize the distances between the image and the text direction sets by matching their first and second moments. Specifically, we align the mean of the image direction set $\mu_{\Delta\mathcal{I}} = \frac{1}{N}\sum_{n=1}^{N}\Delta\mathcal{I}_n$ with the mean of the text direction set $\mu_{\Delta\mathcal{T}} = \frac{1}{K+1}\sum_{i=1}^{K+1}\Delta\mathcal{T}_i$, while matching the covariance of the image direction set $\Sigma_{\Delta\mathcal{I}} = \Delta\mathcal{I}^\top\Delta\mathcal{I}$ with that of the text direction set $\Sigma_{\Delta\mathcal{T}} = \Delta\mathcal{T}^\top\Delta\mathcal{T}$. The directional moment loss is defined as:

$$\mathcal{L}_{dm} = d_1(\mu_{\Delta I}, \mu_{\Delta\mathcal{T}}) + \lambda_{cov} d_2(\Sigma_{\Delta I}, \Sigma_{\Delta\mathcal{T}}), \tag{3}$$

where λ_{cov} is a balancing weighting factor. We instantiate $d_1(\cdot,\cdot)$ with the cosine distance and $d_2(\cdot,\cdot)$ with the euclidean distance. Note that by adding the second term, we can prevent the image directions from being collapsed into a single direction, thus ensuring the sample diversity. A conceptual illustration of our directional moment loss is provided in Figure 2 (b) right.

3.5. Source Knowledge Preservation

To enhance the realism after the adaptation, it is important to preserve valuable content information such as appearances learned from the source domain. For this purpose, StyleGAN-NADA [8] utilizes the layer selection strategy to estimate the importance of each layer and select the top-k important layers to be updated while freezing the rest. However, the number of layers, *i.e.*, k, needs to be tuned for each adaptation scenario, which is cumbersome. Moreover, the frozen layers can also contain valuable information for synthesizing realistic content. In this point of view, we propose to constrain each layer in accordance with its importance for the original task, *i.e.*, source domain generation. To this end, we employ the elastic weight consolidation (EWC) [19] to penalize drastic modification of model parameters. The EWC regularization loss is formulated as:

$$\mathcal{L}_{EWC} = \sum_{l=1}^{L} F^l (\theta_{trg}^l - \theta_{src}^l)^2, \tag{4}$$

where l is the layer index, while θ is the trainable parameters of the generator. The fisher information [30] is estimated as $F = \mathbb{E}\left[-\frac{\partial^2}{\partial \theta_{src}^2} sim(E_I(G_{src}(w)), E_T(t_{src}))\right]$, where $sim(\cdot, \cdot)$ is the cosine similarity between the source text and the generated source samples in the CLIP embedding space.

For more diversity, we design a relation consistency loss that encourages the generator to maintain the semantic relation between images before and after adaptation. Specifically, we extract image features from the generated samples of source and target domain, *i.e.*, $x_{src} = E_I(G_{src}(w)) \in \mathbb{R}^{N \times D}$. With the extracted features, we estimate the interrelation of samples by calculating their dot product similarities $M_{src} = x_{src} \cdot x_{src}^\top \in \mathbb{R}^{N \times N}$, which denotes the inter-sample relation matrix of source samples. The inter-sample relation matrix for target samples M_{trg} can be obtained in a similar way. We apply the row-wise softmax function to M_{src} and M_{trg}, and then minimize the KL divergence to make the target relation similar to the source one. The relation consistency loss is defined as:

$$\mathcal{L}_{rel} = KL(\text{Softmax}(M_{src}), \text{Softmax}(M_{trg})). \tag{5}$$

The overall training objective of the target generator G_{trg} in the second stage, *i.e.*, the adaptation stage, is the weighted sum of the loss functions with two balancing factors λ_{EWC} and λ_{rel} as follows.

$$\mathcal{L}_{S2} = \mathcal{L}_{dm} + \lambda_{EWC} \mathcal{L}_{EWC} + \lambda_{rel} \mathcal{L}_{rel}. \tag{6}$$

4. Experiments

4.1. Implementation Details

Following the setting of the previous work [8], we implement our method based on StyleGANv2 [14] pre-trained on FFHQ [13], AFHQ-Dog, and AFHQ-Cat [5] datasets. The text descriptions of source domains for FFHQ, AFHQ-Dog, and AFHQ-Cat are set to “Photo”, “Dog”, and “Cat”, respectively. In stage 1, the semantic variations $\{z^i\}_{i=1}^K$ are optimized with $\mathcal{L}_{S1}$ during 2,000 iterations. We set the perturbation strength ϵ to the $l2$-norm of the $E_T(t_{trg})$, which is empirically shown to be simple yet effective. The number of semantic variations K is fixed to 6 during evaluation. The balancing weight λ_{div} is set to 1. In stage 2, the generator G_{trg} is trained with the batch size N of 4. For the vision-language model, we employ the pre-trained CLIP [32] with the image encoder of ViT-B/32. The weighting factors λ_{cov}, λ_{EWC}, and λ_{rel} are set to 10^3, 10^7, and 10^2 regarding their loss scales. We utilize the Adam [18] optimizer with the learning rate of 0.002 with betas of $(0, 0.99)$ for variations $\{z^i\}_{i=1}^K$ as well as the generator G_{trg}. We conduct all the experiments on a single RTX 2080Ti GPU.

Table 1. Quantitative results under the “Dog-to-Cat” scenario on AFHQ datasets [5].

Methods	LPIPS (Avg.) (↑)	LPIPS (All) (↑)
Ojha et al. (10-shot) [29]	$0.575_{\pm 0.019}$	$0.575_{\pm 0.046}$
StyleGAN-NADA [8]	$0.460_{\pm 0.010}$	$0.462_{\pm 0.063}$
StyleGAN-NADA [8] + $\mathcal{L}_{EWC}$	$0.480_{\pm 0.006}$	$0.480_{\pm 0.064}$
Baseline ($\mathcal{L}_{dir}$)	$0.402_{\pm 0.008}$	$0.405_{\pm 0.057}$
+ replacing $\mathcal{L}_{dir}$ with $\mathcal{L}_{dm}$	$0.464_{\pm 0.013}$	$0.470_{\pm 0.064}$
+ $\mathcal{L}_{EWC}$	$0.493_{\pm 0.015}$	$0.497_{\pm 0.067}$
+ $\mathcal{L}_{rel}$ (Ours)	$\mathbf{0.507}_{\pm 0.016}$	$\mathbf{0.512}_{\pm 0.072}$

4.2. Quantitative Results

Diversity comparison. Ideally, the generator after adaptation should synthesize the samples of the target domain well while preserving the semantic variations learned from the source domain. To evaluate how well the semantic variations are maintained, we compute the *intra-cluster pairwise LPIPS distance* that directly measures the diversity of generated samples following the existing work [29]. This metric is originally designed for the few-shot setting, where the individual training images are considered to be cluster centroids and the generated samples are clustered using LPIPS [45]. Thereafter, the average LPIPS distance within the cluster is estimated to represent the generated sample diversity. Following StyleGAN-NADA [8], we adapt the metric to the zero-shot setting by building a total of k clusters using k-medoids clustering [15]. For evaluation, we generate 1,000 samples of the target domain for evaluation and compare our method with the state-of-the-art zero-shot method, *i.e.*, StyleGAN-NADA [8], and the few-shot method, *i.e.*, Ojha et al. [29]. We set k to 10 for a fair comparison with Ojha et al. [29] that utilizes 10-shot data.

Table 1 presents the comparison results. The StyleGAN-NADA (second row) denotes the full model equipped with the directional loss $\mathcal{L}_{dir}$ and the layer selection strategy. Noticeably, our method records the average intra-cluster LPIPS score of 0.507, significantly outperforming the existing state-of-the-art zero-shot method, StyleGAN-NADA (0.460). Moreover, our model effectively bridges the gap with the 10-shot method [29] even without using any training examples of the target domain. These results clearly demonstrate the effectiveness of our method in enhancing the diversity of the generated target domain samples.

To better understand where the improvements come from, we break down our method and analyze the effect of each component. Initially, our baseline model equipped with the directional loss $\mathcal{L}_{dir}$ (Eq. 1) shows very poor diversity performance with the average LPIPS of 0.402, indicating that relying on the single target description hinders the model from generating diverse images of the target domain. When augmenting the semantic variations and replacing $\mathcal{L}_{dir}$ with our directional moment loss $\mathcal{L}_{dm}$ (Eq. 3) under our two-stage framework, the diversity of the generated samples is remarkably improved ($0.402 \rightarrow 0.464$), which verifies the importance of modeling the one-to-many relation of the target domain description. On the other hand, the proposed elastic weight consolidation loss $\mathcal{L}_{EWC}$ (Eq. 4) and the relation consistency loss $\mathcal{L}_{rel}$ (Eq. 5) respectively bring additional gains of about 0.029 and 0.014, resulting in the final average LPIPS score of 0.507. This demonstrates the complementarity of the proposed components. To further see the advantage of our elastic weight consolidation loss over the layer selection strategy of StyleGAN-NADA [8], we try applying $\mathcal{L}_{EWC}$ to StyleGAN-NADA instead of the layer selection. As a result, the average LPIPS score is boosted by 0.02, which indicates that suppressing the dramatic changes of important parameters using our $\mathcal{L}_{EWC}$ is more effective than manually selecting the number of layers to be frozen. On the other hand, our full model still surpasses the 'StyleGAN-NADA $+\mathcal{L}_{EWC}$' variant to a large extent, which manifests the importance of exploring semantic variations for sample diversity.

In addition, we also evaluate the diversity with the cluster compactness of generated samples from G_{trg}. We regard generated samples from G_{trg} as a single cluster and estimate the cluster compactness by calculating the sum of squared errors (SSE). To be more specific, we extract features $E_I(G_{trg}(w))$ from the CLIP image encoder, *i.e.*, ViT-B/32, and then compute the cluster center as the mean of image features. SSE is in turn estimated with the sum of squared errors from each feature to the cluster center. In Figure 3 (a), we analyze the change in cluster compactness during 2,000 training iterations, comparing $\mathcal{L}_{dm}$ with $\mathcal{L}_{dir}$. During training, we observe that both losses drop the sample diversity of the source domain during the adaptation, indicating that generating diverse samples is challenging in the zero-shot setting. However, we observe that the slope of $\mathcal{L}_{dm}$ is more gentle compared to $\mathcal{L}_{dir}$, verifying that augmenting semantic variations with the original direction indeed alleviates the mode collapse. Moreover, both $\mathcal{L}_{EWC}$ and $\mathcal{L}_{rel}$ are beneficial for mitigating the collapse and sustaining the sample diversity by effectively preserving important source knowledge and relations.

Additional quantitative evaluations on image fidelity metrics, *i.e.*, FID [10], precision and recall [35, 36, 21], are included in the supplementary material.

User study. To further evaluate the fidelity of generated samples of the target domain, we conduct a user study on the "Cat-to-Dog" adaptation scenario with 58 subjects. We present the generated images from ours and competitors to users to select the best one corresponding to the target domain. As a result, 86.76% of participants favored our results as shown in Figure 3 (b). We also present the participants with 4 images from each method and asked them to choose the one with diverse characteristics. Again, most of the respondents have chosen our method, which indicates that the generated images by ours suffer less from the mode collapse problem and well represent various features of the target domain. The details of the questionnaires for the user study are provided in the supplementary material.

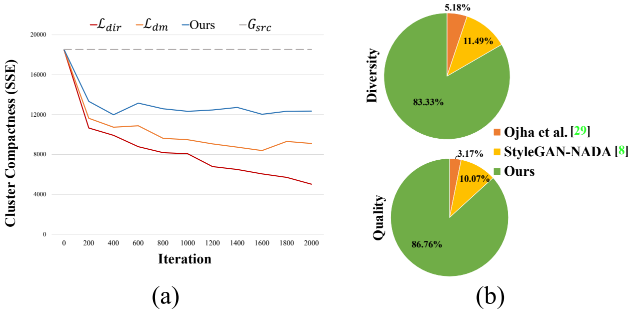


Figure 3. (a) Quantitative comparison of the cluster compactness for diversity evaluation. Higher SSE indicates higher diversity. (b) User study results on the "Cat-to-Dog" scenario. Each participant respond to two questions of preference: 1) quality and 2) diversity of generated images.

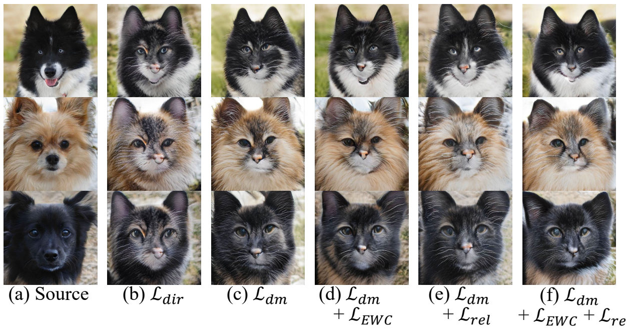

(a) Source
(b) $\mathcal{L}_{dir}$
(c) $\mathcal{L}_{dm}$
(d) $\mathcal{L}_{dm}$ $+\mathcal{L}_{EWC}$
(e) $\mathcal{L}_{dm}$ $+\mathcal{L}_{rel}$
(f) $\mathcal{L}_{dm}$ $+\mathcal{L}_{EWC}+\mathcal{L}_{rel}$

Figure 4. Qualitative ablation study on the "Dog-to-Cat" scenario.

4.3. Ablation Studies.

We conduct ablation studies on our method with the "Dog-to-Cat" scenario qualitatively to verify the effect of each proposed component, whose results are shown in Figure 4. When trained with the directional loss $\mathcal{L}_{dir}$, the generator G_{trg} is guided with only a single direction and loses the sample diversity, synthesizing very similar cat faces with common attributes, *e.g.*, purple ears. On the other hand, $\mathcal{L}_{dm}$ successfully mitigates the problem and helps the generator to generate cat faces with different features, *e.g.*, eyes, ears, and facial directions. In addition with $\mathcal{L}_{EWC}$, quality and diversity is improved with the help of important parameters of the source generator which accounts for naturalness. Furthermore, $\mathcal{L}_{rel}$ emphasizes the characteristics of source images, such as eyes and facial appearances. With the all components combined together, our model produces realistic images of the target domain (cat) while preserving

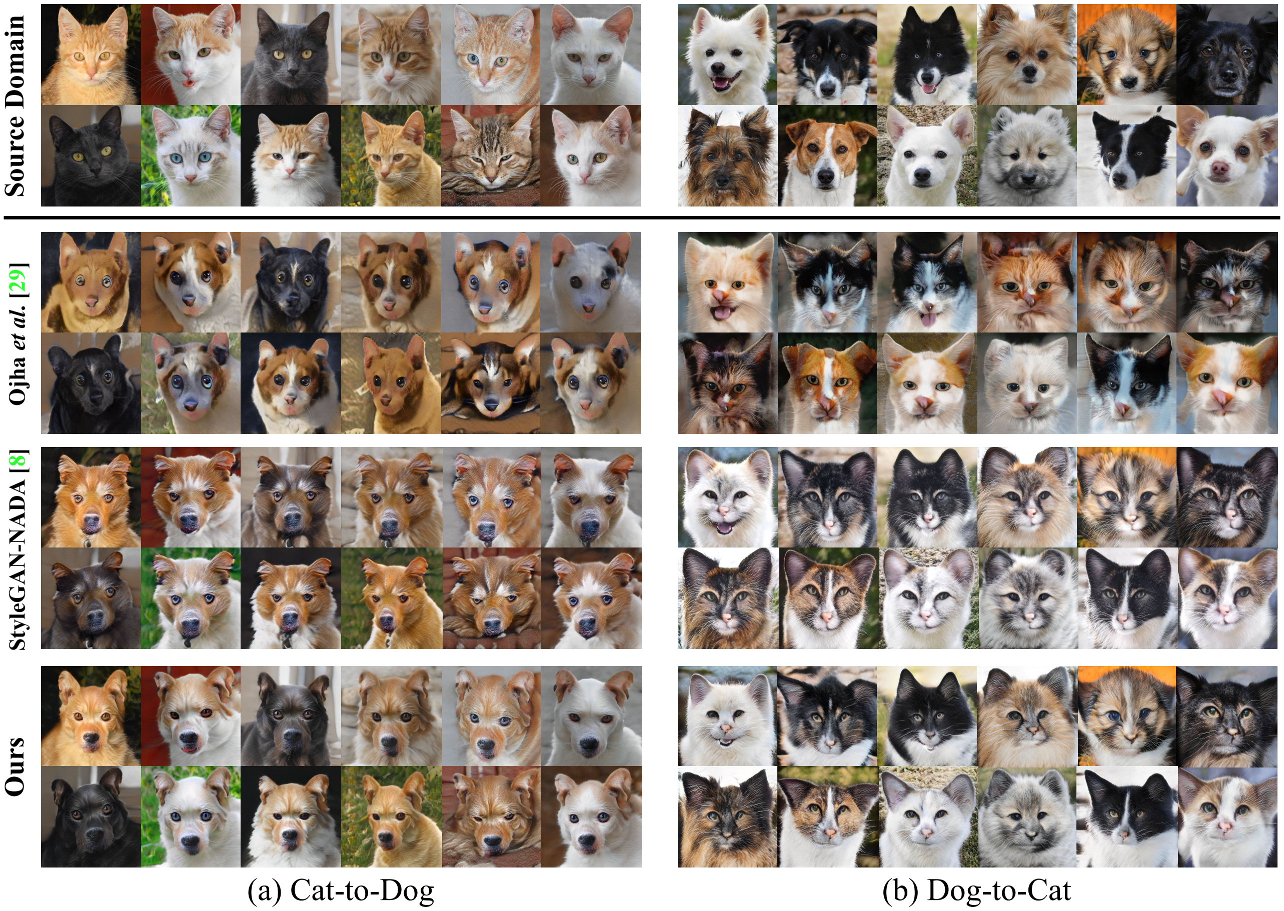


Figure 5. Qualitative comparison on AFHQ datasets [5]. We compare our model to the previous state-of-the-art zero-shot [8] and few-shot [29] methods under the two scenarios: (a) "Cat-to-Dog" and (b) "Dog-to-Cat". Here Ojha et al. [29] is trained in the 10-shot setting.

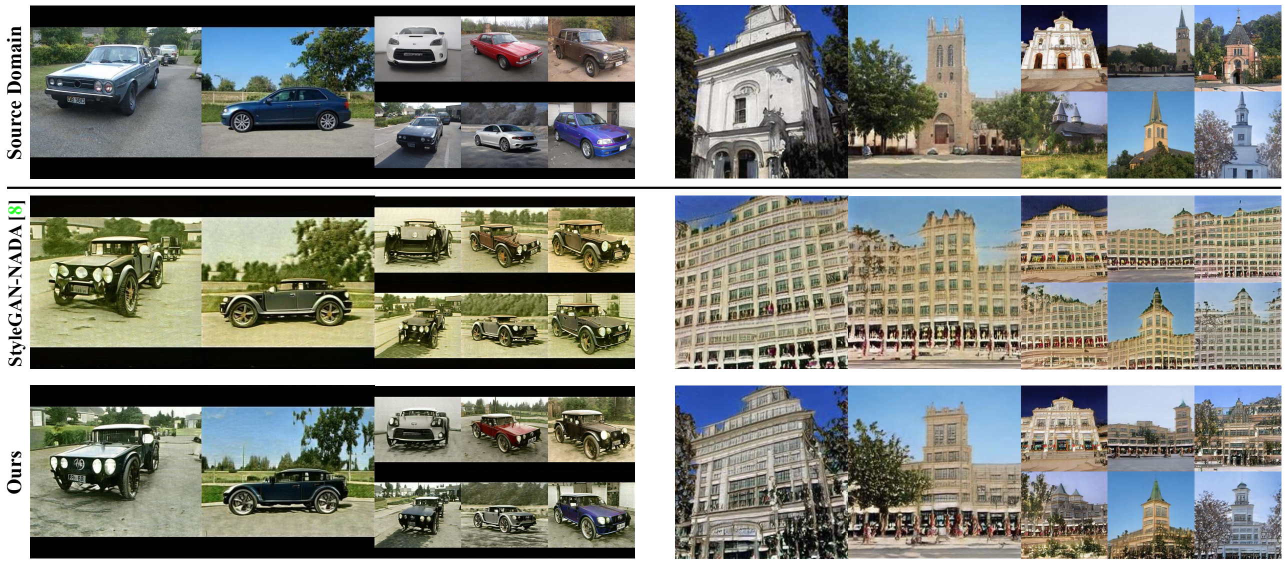
diverse diverse attributes of the source domain (dog).

4.4. Qualitative Results

In Figure 5, we compare our method qualitatively with StyleGAN-NADA [8] and Ojha et al. [29] on "Cat-to-Dog" and "Dog-to-Cat" scenarios. Note that we randomly sample 10 training images from the AFHQ dataset to train the few-shot GAN adaptation method of Ojha et al. We present the generated samples from the same latent codes for fair comparisons. As shown in Figure 5 (a), the generated samples from StyleGAN-NADA largely lose the diversity, showing dog images that resemble each other. Consistently, the generated cats by StyleGAN-NADA in Figure 5 (b) show very similar facial characteristics and expressions without discrimination. Meanwhile, Ojha et al. preserve diversity but fail to achieve high quality in both scenarios. In contrast, our method successfully generates realistic images with diverse characteristics of the target domain.

Also in Figure 6, we display the qualitative results with object adaptation scenarios, *i.e.*, "Car-to-Car in 1920s" and "Church-to-Department Store", using the source generator trained on LSUN [42] dataset. Since StyleGAN-NADA heavily depends on the single target text feature, the results lack the diversity while reflecting the common design. For example, the diverse characteristics of the cars in the source domain, *e.g.*, shapes and colors, are diminished after adaptation (Figure 6 (a)). In addition, the generated department stores all have same repetitive windows, while the original contexts are collapsed with the entire image filled with the building. In contrast, our model synthesizes more natural images of the target objects with different designs reflecting the source contextual variations. More qualitative results and comparisons in various adaptation scenarios are provided in the supplementary material.

To demonstrate that the proposed framework can also be utilized for text-guided image editing, we show manipulation results of real images with the text prompts. We employ StyleGAN [13] pretrained on FFHQ as the source generator and sample the images from the CelebA test split [24] as the manipulation target. To embed the real images into the latent space, we exploit GAN inversion methods [33, 2, 1]. Afterwards, we feed-forward the obtained latent codes to



(a) Source : Car / Target : Car in 1920s (b) Source : Church / Target : Department Store

Figure 6. Qualitative comparison in the object translation scenarios. The source generators are pre-trained on LSUN-Car and LSUN-Church datasets [13] and adapted to the target text descriptions ”Car in 1920s” and ”Department Store”, respectively. Noticeably, the proposed framework enhances both the diversity and quality of the generated samples.

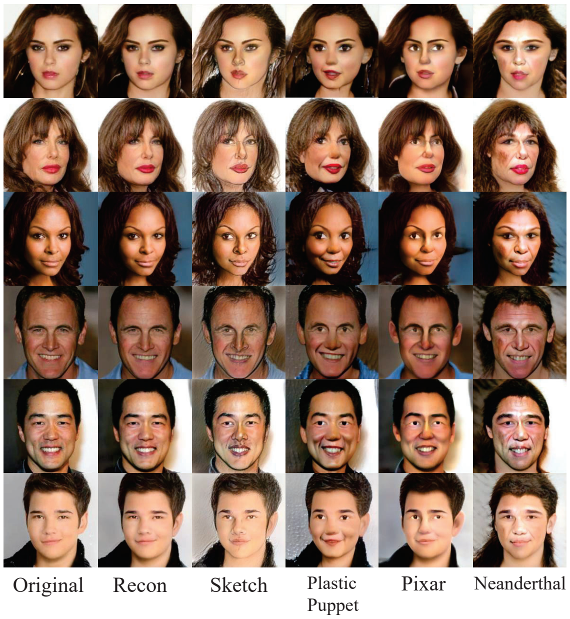


Figure 7. Image manipulation results on different target domains. The real images are sampled from CelebA [24] dataset and inverted into latent space via psp [33] trained on FFHQ [13].

the target generator adapted to the designated text to get the final results. Since the mapping network and the latent space remain unchanged, the source and target images from the same latent code share the same identity. On the other hand, the parameters of the target generator are updated to align the image editing direction with the text-guided direction from the source to the target domain. As shown in Figure 7, our framework successfully translates the real images into various target domains while preserving the personal characteristics.

5. Conclusion

In this paper, we proposed a novel zero-shot GAN adaptation framework that can generate diverse samples of the target domain. Specifically, we introduced a novel method to find semantic variations of the target text in CLIP embedding space and propose a directional moment loss for encouraging the target generator to learn the diverse characteristics of the target domain. Furthermore, in order to preserve the knowledge obtained from the source domain, we employ elastic weight consolidation (EWC) to regularize the drastic parameter updates of the generator. In addition, we introduce a relation consistency loss for more diversity. Through experiments on various adaptation scenarios, we demonstrate that our proposed methods ensure the target sample diversity both qualitatively and quantitatively. In addition, our model achieves a new state-of-the-art on the task of zero-shot GAN adaptation.

Acknowledgements. This research was partly supported by the MSIT (Ministry of Science, ICT), Korea, under the High-Potential Individuals Global Training Program (No. 2021-0-01696) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), and the National Research Foundation of Korea grant funded by the Korean government (MSIT) (No. 2022R1A2B5B02001467). This project is supported by Microsoft Research Asia.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021.

[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] Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. Data-efficient language-supervised zero-shot learning with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3119–3124, 2021.

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Aviv Gabbay, Niv Cohen, and Yedid Hoshen. An image is worth more than a thousand words: Towards disentanglement in the wild. *Advances in Neural Information Processing Systems*, 34, 2021.

[8] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

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

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[11] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.

[12] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[15] Leonard Kaufman and Peter J Rousseeuw. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 344:68–125, 1990.

[16] Gwanghyun Kim and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. *arXiv preprint arXiv:2110.02711*, 2021.

[17] Seongtae Kim, Kyoungkook Kang, Geonung Kim, Seung-Hwan Baek, and Sunghyun Cho. Dynagan: Dynamic few-shot adaptation of gans to multiple domains. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022.

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

[19] 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.

[20] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. *arXiv preprint arXiv:2112.00374*, 2021.

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

[22] Yijun Li, Richard Zhang, Jingwan (Cynthia) Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15885–15896. Curran Associates, Inc., 2020.

[23] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*, 2020.

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[25] Mario Lučić, Michael Tschannen, Marvin Ritter, Xiaohua Zhai, Olivier Bachem, and Sylvain Gelly. High-fidelity image generation with fewer labels. In *International conference on machine learning*, pages 4183–4192. PMLR, 2019.

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

[27] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in Neural Information Processing Systems*, 34, 2021.

[28] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2750–2758, 2019.

[29] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021.

[30] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*, 2013.

[31] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[33] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021.

[34] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *arXiv preprint arXiv:2010.11943*, 2020.

[35] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in n eural information processing systems*, 31, 2018.

[36] Loic Simon, Ryan Webster, and Julien Rabin. Revisiting precision recall definition for generative modeling. In *International Conference on Machine Learning*, pages 5799–5808. PMLR, 2019.

[37] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. Towards good practices for data augmentation in gan training. *arXiv preprint arXiv:2006.05338*, 2:3, 2020.

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

[39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020.

[40] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 218–234, 2018.

[41] Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. Few shot generative model adaption via relaxed spatial structural alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11204–11213, 2022.

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

[43] Nir Zabari and Yedid Hoshen. Semantic segmentation in-the-wild without seeing any segmentation examples. *arXiv preprint arXiv:2112.03185*, 2021.

[44] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations*, 2019.

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[46] Zicheng Zhang, Yinglu Liu, Congying Han, Tiande Guo, Ting Yao, and Tao Mei. Generalized one-shot domain adaption of generative adversarial networks. *Advances in Neural Information processing systems*, 2022.

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

[48] Yunqing Zhao, Henghui Ding, Houjing Huang, and Ngai-Man Cheung. A closer look at few-shot image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9140–9150, 2022.

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

[50] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. *arXiv preprint arXiv:2110.08398*, 2021.