Zero-shot Generative Model Adaptation via Image-specific Prompt Learning

Jiayi Guo[1*] Chaofei Wang[1*] You Wu[2] Eric Zhang[3] Kai Wang[3] Xingqian Xu[3] Shiji Song[1]
Humphrey Shi[3,4†] Gao Huang[1†]

[1]Tsinghua University, BNRist [2]UCAS [3]SHI Labs @ Oregon & UIUC [4]Picsart AI Research (PAIR)

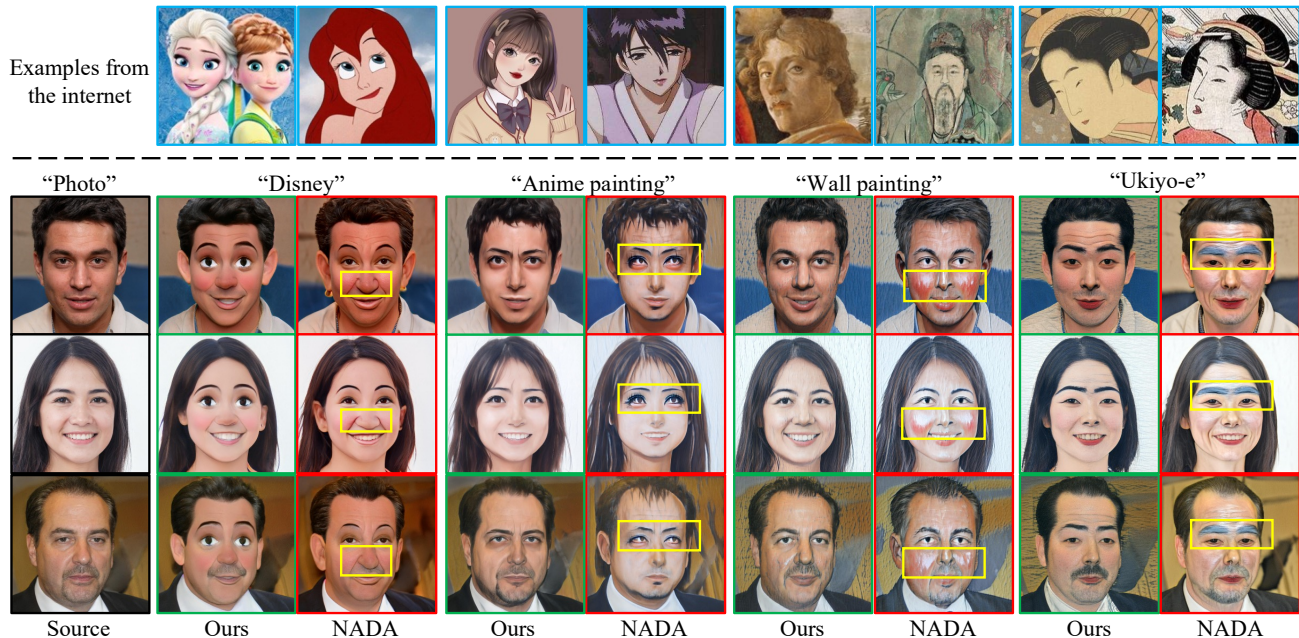https://github.com/Picsart-AI-Research/IPL-Zero-Shot-Generative-Model-Adaptation

Figure 1. The mode collapse issue. For NADA [21] and our method, the same generator pre-trained on the source domain of "Photo" is adapted to the unseen target domains of "Disney", "Anime painting", "Wall painting" and "Ukiyo-e" only with the domain labels. The images above the dotted line are some examples from the internet. The generated images of NADA exhibit some similar unseen patterns (yellow box areas) which are undesired in terms of quality and diversity. This issue is largely addressed by our method.

Abstract

*Recently, CLIP-guided image synthesis has shown appealing performance on adapting a pre-trained source-domain generator to an unseen target domain. It does not require any target-domain samples but only the textual domain labels. The training is highly efficient, e.g., a few minutes. However, existing methods still have some limitations in the quality of generated images and may suffer from the mode collapse issue. A key reason is that a fixed adaptation direction is applied for all cross-domain image pairs, which leads to identical supervision signals. To address this issue, we propose an **Image-specific Prompt Learning (IPL)** method, which learns specific prompt vectors for each source-domain image. This produces a more precise adaptation direction for every cross-domain image pair, endowing the target-domain generator with greatly enhanced flexibility. Qualitative and quantitative evaluations on various domains demonstrate that IPL effectively improves the quality and diversity of synthesized images and alleviates the mode collapse. Moreover, IPL is independent of the structure of the generative model, such as generative adversarial networks or diffusion models. Code is available at https://github.com/Picsart-AI-Research/IPL-Zero-Shot-Generative-Model-Adaptation.*

1. Introduction

In recent years, image synthesis using generative adversarial networks (GANs) [11] has been rapidly developed.

---

*Equal contribution.

†Corresponding authors.

The state-of-the-art methods can generate images that are hard to be distinguished from real data [14, 20, 21, 46, 50]. However, the GAN-based methods heavily rely on *vast quantities* of training examples, and adopt a cumbersome adversarial training scheme which generally costs many hours of training time. Unfortunately, in many real-world scenarios, data acquisition is difficult or expensive. For example, in the artistic domains, it is impossible to have artists make thousands of creations. The high training cost is also unacceptable on some embedded devices, e.g., cellphones.

To address these issues, researchers begin to focus on the generative model adaptation. The goal of this task is to adapt a pre-trained source-domain generator to a target domain with *limited* data. Many few-shot GAN-based methods are proposed, such as TGAN [48], FreezeD [30], MinGAN [47], ADA [18], DiffAug [53], IDC [33] and RSSA [49], etc. However, these methods still require some training images of the target domain and follow the adversarial training scheme. As a pioneer work, StyleGAN-NADA [8] (NADA for short) proposes a *zero-shot* adaptation method, which only requires textual domain labels and discards the cumbersome adversarial training scheme by introducing a pre-trained CLIP model. Although efficient, it still has obvious deficiencies, i.e., the limited quality and mode collapse of generated images. As shown in Fig.1, we adapt a pre-trained generator of "Photo" domain to "Disney", "Anime painting", "Wall painting" and "Ukiyo-e" domains. For the results of NADA [8], we notice that the generated images of the same target domain always show some homogeneous patterns which degrade the image quality and diversity, such as deep nasolabial folds in "Disney", squinting eyes in "Anime painting", red cheeks in "Wall painting" and blue eyebrows in "Ukiyo-e" (yellow box areas).

By exploring the factors behind this phenomenon, we find that the key factor is the *fixed* adaptation direction produced by manually designed prompts. Sharing the direction for all cross-domain image pairs leads to identical supervision signals for the model adaptation. Consider the example, adapting a generator of "Human" domain to "Tolkien elf" domain as shown in Fig.2. The previous works [8, 22] adopt manually designed prompts (e.g., "A photo of a") plus the domain label to produce a fixed adaptation direction, which is *shared* by all cross-domain image pairs (Fig.2 (a)) in the adaptation process. We argue that the constraint is too restrictive and suppresses the image-specific features, leading to homogeneous generated patterns.

In this paper, we propose an **I**mage-specific **P**rompt **L**earning (IPL) method to address the above issue. The motivation is setting more precise and diversified adaptation directions by customizing more image-specific prompts, for instance "Asian girl", "Curly hair lady" and "Elder glass man" (Fig.2 (b)). These adaptation directions endow the target-domain generator with high flexibility to synthesize

(a) Manual prompts: a shared fixed reference direction

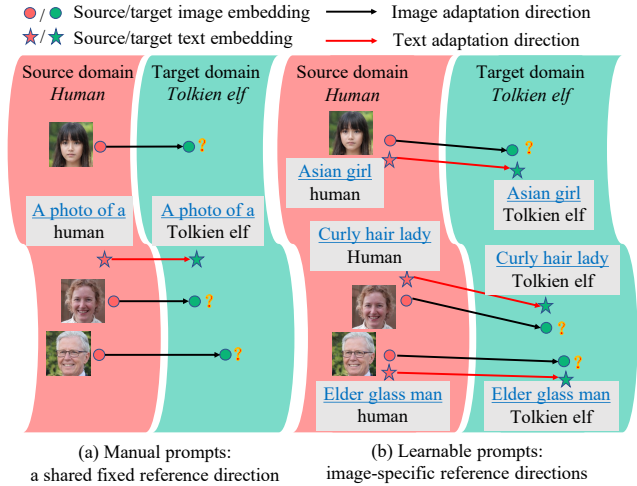(b) Learnable prompts: image-specific reference directions

Figure 2. An illustration of our motivation. The previous methods adopt manual prompts to compute a fixed adaptation direction for all cross-domain image pairs, while our method learns image-specific prompts for producing more precise and diversified adaptation directions.

more diversified images. The proposed IPL is a two-stage method. In Stage 1, a latent mapper is trained to produce an *image-specific* set of prompt vectors conditioned on each source-domain image by a contrastive training scheme. The learned prompt vectors contain more specific and diversified features of the source-domain images than the fixed prompt vectors. We further propose a domain regularization loss to ensure that the learned prompt vectors are compatible with the target domain. In Stage 2, we compute more *precise* and *diversified* adaptation directions for each cross-domain image pair, and train the target-domain generator with an adaptive directional CLIP loss, which can be viewed as an improved version of the Directional CLIP Loss [8]. As shown in Fig.1, our method alleviates the mode collapse issue well. Extensive experiments across a wide range of domains demonstrate that the proposed IPL effectively improves the quality of synthesized images and overcomes the mode collapse issue. User studies and ablation studies are also conducted to validate the effectiveness of our method.

It is worth noting that our proposed IPL method is independent of the structure of the generative model, and can be applied to the recent diffusion models [13, 27, 31, 35, 41–43, 51]. Thus we also combine IPL with diffusion models and get a more robust and stronger generative capacity, especially on complex images, which shows the high effectiveness and adaptability of our approach.

2. Related Work

Generative model adaptation. Generative model adaptation is the task of adapting a generative model trained on a large-scale source domain to a data-limited target domain.

According to the size of the training dataset of the target domain, it can be directly divided into two main categories: few-shot generative model adaptation and zero-shot generative model adaptation. For the few-shot generative model adaptation task, the most natural approach is to fine-tune a pre-trained GAN [2, 4, 26, 48]. However, fine-tuning the entire network weights used to result in overfitting. Subsequently, many methods were proposed to alleviate the overfitting issue. They either imposed strong regularization [52, 54], or modified the network parameters with a slight perturbation [30, 32, 37, 47], or preserved some important information by cross-domain alignment [33, 49], or performed data augmentation [45, 53, 55].

For the zero-shot generative model adaptation task, NADA [8] first proposed to introduce a pre-trained CLIP model for supplying necessary prior knowledge. It only required textual domain labels, and encoded the domain gap as a text-guided adaptation direction in CLIP space. To enhance the identity-preserving capability of real-world image translation, Kim *et al*. further proposed DiffusionCLIP [22] which utilized diffusion models [42] instead of StyleGANs [18–21] in NADA. Nevertheless, these existing works all adopt a fixed adaptation direction which only contains the basic domain knowledge but no image-specific features. In this paper, we argue that this shared fixed adaptation direction may lead to the mode collapse issue. To produce more accurate and adaptive adaptation directions, we propose to learn diverse and specific prompt vectors for each image.

Prompt learning. Prompt engineering is first introduced as a knowledge probing approach [34]. Given cloze-style prompts, it induces pre-trained language models to generate the corresponding answers. However, manually designed prompts may be sub-optimal and provide imprecise guidance. To tackle this issue, prompt learning [9, 16, 23, 25, 28, 40, 56] has been widely studied in natural language processing to automatically explore the optimal set of prompts. With the unprecedented development of vision-language models [15, 36] in recent years, researchers begin to apply prompt learning to computer vision tasks [7, 10, 17, 24, 57, 58]. In specific, Zhou *et al*. [57, 58] first adopted context optimization in image classification tasks by modeling context words with continuous vectors in the word embedding space. Subsequently, many downstream tasks in computer vision were also explored, e.g., object detection [7], visual grounding [24], video understanding [17] and transfer learning [10]. As far as we know, this is the first work to propose an adaptive prompt learning scheme for generative model adaptation. Different from previous prompt learning schemes, our method introduces a latent mapper to learn a specific set of prompt vectors for each image. When training the target-domain generator, the learned image-specific prompt vectors could produce more precise adaptation directions to provide better supervision signals.

3. Methodology

The goal of zero-shot generative model adaptation is to adapt a pre-trained source-domain generator G_s to an unseen target domain, and get the target-domain generator G_t. The source domain with the domain label Y_s, e.g., "Human", can obtain plentiful high-quality images by G_s. The target domain is described only through the domain label Y_t, e.g., "Tolkien elf", with no images. Following [8, 22], a pre-trained CLIP model [36] including an image encoder E_I and a text encoder E_T is introduced.

We propose a two-stage method named **I**mage-specific **P**rompt **L**earning (IPL). Its framework is shown in Fig.3. In Stage 1, a latent mapper F is trained to produce a set of image-specific prompt vectors $\{[\mathbf{V}]_1^i, [\mathbf{V}]_2^i, \cdots, [\mathbf{V}]_m^i\}$ for each latent code w^i of a source-domain image. Each prompt vector has the same dimension with word embeddings in CLIP space. The training loss consists of a contrastive learning loss $\mathcal{L}_{\text{contr}}$ and a domain regularization loss $\mathcal{L}_{\text{domain}}$. The former aims to preserve the image-specific features of each source domain image in the learned prompt vectors. The latter constrains the image-specific features to be suitable to the target domain, which means the learned features should not conflict with the target domain. For example, the features of prompts like "round ear" should not be contained in the ideal prompt vectors if the target domain is "Tolkien elf". In Stage 2, the trained latent mapper F is plugged into the training process of the target-domain generator G_t, and produces more precise and diversified adaptation directions for cross-domain image pairs. This training stage follows [8] except that learned prompt vectors produced by the latent mapper F replace the fixed prompt vectors. The final textual supervision information includes shared learned prompt vectors and respective embeddings of the original domain labels.

3.1. Image-specific prompt learning

General prompts. The previous methods [8, 22] compute a fixed adaptation direction produced by two embeddings of manually designed prompts, e.g., "a photo of a human" and "a photo of a Tolkien elf", then constrain the directions of all cross-domain pairs to be parallel with the adaptation direction. In contrast to manually designed prompts, prompt learning [58] aims to find the optimal set of prompt vectors for a domain by directly tuning the embeddings of prompts. Formally, we define a general prompt matrix M_d to represent a given domain d. M_d consists of the prompt vectors $[\mathbf{V}]_1, [\mathbf{V}]_2, \cdots, [\mathbf{V}]_m$ and the embedding of the domain label $[Y_d]$ as below:

$$M_d = [\mathbf{V}]_1 [\mathbf{V}]_2 \cdots [\mathbf{V}]_m [Y_d], \tag{1}$$

where m is the number of prompts. Suppose the dimension of each embedding is k. Then the dimension of M_d should be $(m + 1) \times k$. In [8, 22], the prompt vectors

Stage 1: Training latent mapper for prompt learning
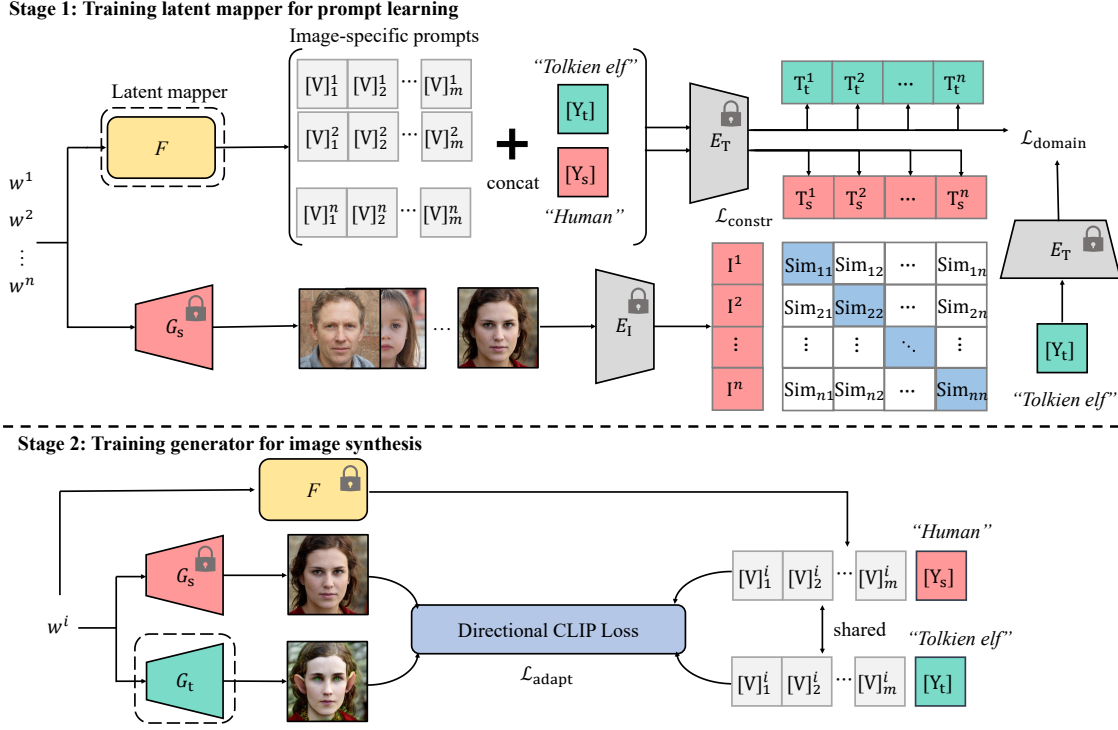
Stage 2: Training generator for image synthesis

Figure 3. The framework of our method. In Stage 1, a latent mapper F is trained for prompt learning by a contrastive learning loss $\mathcal{L}_{\text{contr}}$ and a domain regularization loss $\mathcal{L}_{\text{domain}}$. The image encoder E_{I} and the text encoder E_{T} are from the CLIP model [36]. In Stage 2, the target-domain generator G_{t} is trained for image synthesis by the improved Directional CLIP Loss $\mathcal{L}_{\text{adapt}}$ in which the adaptive prompts produced by the latent mapper are applied. In two stages, the locked modules are fixed while the unlocked modules are trained. For simplicity, we replace $E_{\text{I}}(G_{\text{s}}(w^i))$ and $E_{\text{T}}(\text{M}_{\text{s}}^i)$ with I^i and T_{s}^i, respectively.

$[\mathbf{V}]_1, [\mathbf{V}]_2, \cdots, [\mathbf{V}]_m$ are fixed embeddings of manually designed prompts. For prompt learning [58], the prompt vectors are learned by encoding each training image of the domain d with E_{I} and the prompt matrix M_{d} with E_{T}, and then maximizing the cosine similarity between them.

Inspired by prompt learning, in the zero-shot generative model adaptation task, a natural idea is to learn an optimal set of prompt vectors instead of the manually designed prompts in NADA [8]. Although the adaptation direction calculated by the learned prompt vectors seems to be more reasonable than that of the manually designed prompts, it is still fixed and shared for all cross-domain image pairs. These fixed learned prompt vectors can not solve the mode collapse issue (Experimental validations can be seen in Sec. 4.4). To obtain more flexible and diversified adaptation directions, we further propose to learn a set of image-specific prompt vectors for each image, which can be regarded as an improved version of prompt learning.

Image-specific prompts. Utilizing the source-domain generator G_{s}, we train a latent mapper F as shown in Fig.3 (Stage 1). Through the mapper, each image of the source domain can be matched to an optimal set of prompt vectors. Formally, given a latent code w^i, corresponding to

the ith image in the source domain, the image-specific set of prompt vectors $\{[\mathbf{V}]_1^i, [\mathbf{V}]_2^i, \cdots, [\mathbf{V}]_m^i\}$ can be obtained by $F(w^i, \theta)$, where θ denotes the parameters of the latent mapper F. Following the definition of the prompt matrix in Eq.(1), we define an image-specific prompt matrix of the ith source-domain image as:

$$\text{M}_{\text{s}}^i = F(w^i, \theta)[\text{Y}_{\text{s}}] = [\mathbf{V}]_1^i[\mathbf{V}]_2^i \cdots [\mathbf{V}]_m^i[\text{Y}_{\text{s}}]. \quad (2)$$

In this paper, F is a common four-layer fully-connected network. Next, we show how to train it.

Contrastive training scheme. Given a batch of latent codes $\{w^1, w^2, ..., w^n\}$, we can produce a batch of sets of prompt matrices $\{\text{M}_{\text{s}}^1, \text{M}_{\text{s}}^2, ..., \text{M}_{\text{s}}^n\}$ by F and a batch of images $\{G_{\text{s}}(w^1), G_{\text{s}}(w^2), ..., G_{\text{s}}(w^n)\}$ by G_{s}. Then $n \times n$ pairs $< G_{\text{s}}(w^i), \text{M}_{\text{s}}^j >, i, j \in \{1, 2, ..., n\}$ have been obtained. Then, we take the pairs of $i = j$ as positive samples, and the pairs of $i \neq j$ as negative samples for contrastive training. Specifically, we compute the similarity between embeddings of the ith image and the jth prompt matrix in CLIP space as:

$$\text{Sim}_{ij} = \text{Cos}(\text{Norm}(E_{\text{I}}(G_{\text{s}}(w^i))), \text{Norm}(E_{\text{T}}(\text{M}_{\text{s}}^j))), \quad (3)$$

where $\text{Norm}(\cdot)$ and $\text{Cos}(\cdot)$ represent L_2 normalization and the cosine function, respectively. The similarities of pos-

itive samples are maximized while the similarities of negative samples are minimized. The contrastive loss is expressed as:

$$\mathcal{L}_{\text{contr}} = \mathbb{E}_{w \in \mathcal{W}}(\sum_{i \neq j}(\text{Sim}_{ij}) - \sum_{i=j}(\text{Sim}_{ij})). \quad (4)$$

Domain regularization loss. For the target domain without any prior knowledge except the domain label Y_t, we can simply share the learned prompt vectors between the source and target domains following [8]. However, the shared prompt vectors may lead to the risk of generating unrealistic images for the target domain, because some learned prompt vectors may contain strongly relevant features to the source domain, leading to conflict with the target domain. For example, an image of "Human" domain is matched to prompt vectors of "round ear", but a corresponding image of "Tolkien elf" domain should not contain the features of "round ear". Sharing these prompt vectors is harmful to the target-domain image generation. Therefore, we further propose a domain regularization loss. Specifically, we constrain the angles between the embeddings of the image-specific prompt matrix M_t^i and the target-domain label Y_t in CLIP space to be small, to avoid the learned prompt vectors conflicting with the target domain. Formally, the domain regularization loss is described as:

$$\mathcal{L}_{\text{domain}} = -\mathbb{E}_{w^i \in \mathcal{W}} \sum_{i=1}^{n}(\text{Cos}(E_T(M_t^i), E_T(Y_t))), \quad (5)$$

where M_t^i is calculated by Eq.(2) except replacing the domain label, $\text{Cos}(\cdot)$ represents the cosine similarity.

As a summary, the whole training loss function of the latent mapper F is:

$$\mathcal{L} = \mathcal{L}_{\text{constr}} + \lambda \mathcal{L}_{\text{domain}}, \quad (6)$$

where λ is the ratio parameter. Optimized by $\mathcal{L}$, the learned prompt vectors can not only reflect the features of the source-domain images, but also adapt to the target domain.

3.2. Latent mapper guided generator training

After training the latent mapper F, we conduct the second stage: training the target-domain generator G_t as shown in Fig.3 (Stage 2). In specific, we plug in the trained latent mapper, and train G_t with an improved Directional CLIP Loss $\mathcal{L}_{\text{adapt}}$. Its main difference with [8] is using the image-specific prompt vectors that are produced on-the-fly by F instead of the fixed ones of manually designed prompts. Formally, given a latent code w^i, we calculate the direction of the ith source and target image pair as below:

$$\Delta I_i = \text{Norm}(E_I(G_t(w^i))) - \text{Norm}(E_I(G_s(w^i))), \quad (7)$$

where $\text{Norm}(\cdot)$ represents L_2 normalization. The image-specific adaptation direction is calculated as below:

$$\Delta T_i = \text{Norm}(E_T(M_t^i)) - \text{Norm}(E_T(M_s^i)). \quad (8)$$

The improved Directional CLIP Loss $\mathcal{L}_{\text{adapt}}$ is:

$$\mathcal{L}_{\text{adapt}} = \mathbb{E}_{w^i \in \mathcal{W}} \sum_{i=1}^{n}(1 - \frac{\Delta I_i \cdot \Delta T_i}{|\Delta I_i||\Delta T_i|}), \quad (9)$$

where n is the batch size of latent codes. $\mathcal{L}_{\text{adapt}}$ constrains the direction of each image pair ΔI_i with an image-specific adaptation direction ΔT_i.

4. Experiments

In this section, we evaluate our method qualitatively and quantitatively. The experimental setup is firstly presented in Sec. 4.1. Then we show image synthesis results across various domains in Sec. 4.2. Utilizing a GAN inversion model and diffusion models, results of real-world image translation are provided in Sec. 4.3. Finally, we carefully conduct ablation studies on prompt designing schemes and loss term ratios in Sec. 4.4.

4.1. Experimental setup

Baselines and settings. Two strong methods are chosen as our competitors. For zero-shot image synthesis, NADA [8] is the state-of-the-art method. Following NADA [8], we adapt the pre-trained StyleGANv2 [21] generators on (i) Flickr-Faces-HQ (FFHQ) [8] and (ii) Animal FacesHQ (AFHQ) [3], utilize the same pre-trained CLIP [36] built on ViT-B/32 [6]. For zero-shot real-world image translation, we utilize Restyle [1] with e4e [44] encoder to invert a real image into the latent space $\mathcal{W}$ for StyleGANs. DiffusionCLIP (Diff-CLIP for short) [22] is the state-of-the-art method. We follow the setting of [22] except replacing denoising diffusion implicit models (DDIM) [42] with diffusion autoencoders [35]. The training process includes 300 iterations for prompt learning and 300 iterations for generator adaptation using a single NVIDIA RTX 3090 GPU. The batch size is set to 32 for prompt learning and 2 for generator adaptation. The number of learned prompt vectors m is set to 4. For each domain, the ratio parameter λ in Eq.(6) is selected among [1, 10], according to the best Inception Score [38] of adapted generators. The whole training process requires about 10~20 minutes. More implementation details can be seen in supplementary materials.

Evaluation metrics. The ideal generated images should have: 1) high quality and diversity, 2) correct target-domain style, and 3) necessary source-domain information preservation (e.g., structure or identity). For a comprehensive evaluation, we utilize the popular Inception Score (IS) [38] to evaluate the image quality and diversity, the Single Image Fréchet Inception Distance (SIFID) [39] to evaluate the target-domain style, the Structural Consistency Score (SCS) [49] to evaluate the structure preservation, the identity similarity (ID) [5,12] to evaluate the identity preservation. More details can be seen in supplementary materials.
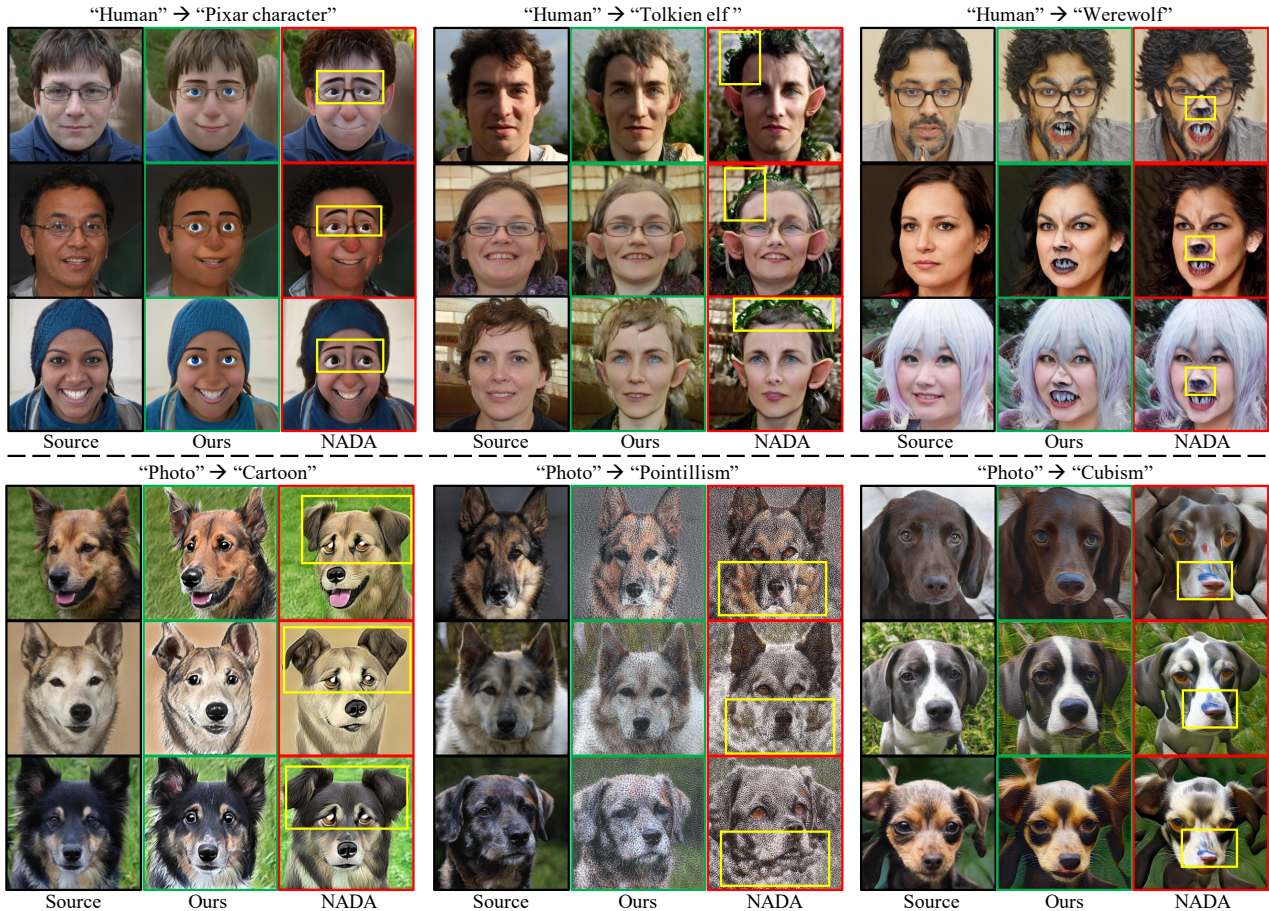
Figure 4. Image synthesis comparison results. For FFHQ [21], the source domain is "Human" and the target domains are "Pixar character", "Tolkien elf", and "Werewolf". For AFHQ-Dog [3], the source domain is "Photo" and the target domains are "Cartoon", "Pointillism", and "Cubism". The yellow box areas show the mode collapse problem of NADA [8].

Table 1. Quantitative evaluation results. US denotes user study. The best results are **bold**.

Dataset	Source→Target	IS [38] (↑)		SCS [49] (↑)		ID [5,12] (↑)		SIFID [39] (↓)						US (↑)
		NADA	IPL	NADA	IPL	NADA	IPL	NADA			IPL			
								R_1	R_2	R_3	R_1	R_2	R_3	
	Photo→Disney	2.721	**3.089**	0.407	**0.448**	0.782	**0.801**	2.776	3.136	3.670	**2.517**	**2.930**	**3.497**	82.6%
	Photo→Anime painting	2.450	**3.051**	0.324	**0.518**	0.666	**0.776**	2.956	1.811	1.242	**2.845**	**1.595**	**1.021**	79.3%
	Photo→Wall painting	2.183	**2.676**	0.439	**0.487**	0.594	**0.637**	1.944	1.220	1.331	**1.930**	**1.183**	**1.274**	80.9%
FFHQ [8]	Photo→Ukiyo-e	2.205	**2.974**	0.420	**0.506**	**0.775**	0.632	1.954	1.990	1.326	**1.165**	**1.255**	**0.878**	85.9%
	Human→Pixar character	2.703	**2.785**	0.379	**0.461**	0.757	**0.853**	0.793	0.932	**0.865**	0.638	**0.821**	1.092	86.7%
	Human→Tolkien elf	2.479	**2.778**	0.416	**0.491**	0.711	**0.772**	**0.632**	1.495	1.452	0.690	**0.637**	**0.701**	76.8%
	Human→Werewolf	2.619	**2.809**	0.399	**0.417**	0.642	**0.747**	1.969	1.846	1.967	**1.734**	**1.688**	1.911	72.7%
AFHQ [3]	Photo→Cartoon	6.505	**8.658**	0.407	**0.563**	0.925	**0.941**	2.708	2.672	3.870	**2.517**	**2.477**	**3.278**	87.6%
	Photo→Pointillism	5.419	**6.913**	0.224	**0.542**	0.775	**0.881**	7.081	5.288	7.142	**4.818**	**3.089**	**4.074**	78.5%
	Photo→Cubism	4.165	**6.450**	0.386	**0.463**	0.934	**0.943**	2.779	**2.938**	3.199	**2.431**	2.956	**2.284**	74.3%

4.2. Generative model adaptation

Qualitative comparison. In addition to Fig.1, we conduct extensive experiments across a wide range of domains as shown in Fig.4. All results indicate that our proposed approach outperforms NADA consistently. The yellow box areas in the figures denote the main different features between NADA and our IPL. From the quality of the gener-

ated images, the results of NADA have more incorrect features and noise, such as green mussy noise on hairs (Tolkien elf), ruined noses (Werewolf) and unshaped necks (Pointillism), while the results of IPL are more clear and correct. From the mode collapse perspective of the generated images, NADA is prone to collapse to some similar facial features for different images, such as depressed emotions (Pixar character), folded ears (Cartoon) and blue noses (Cu-
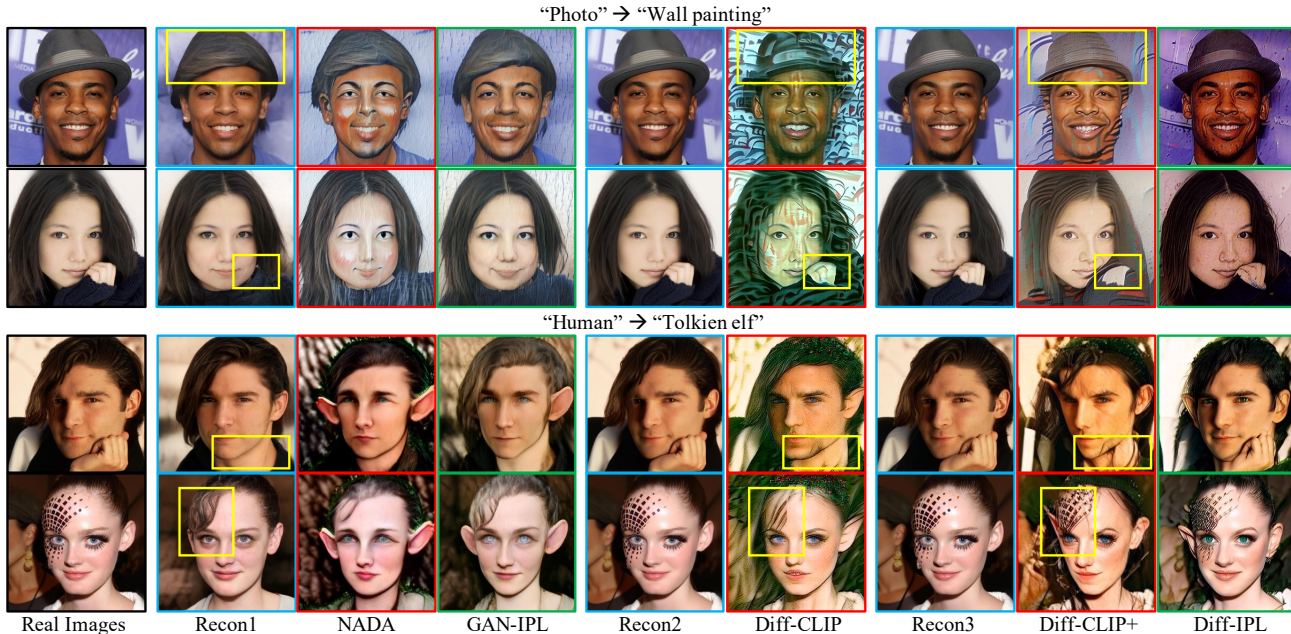
Figure 5. Real-world image translation comparison results. Baselines are NADA [8], Diff-CLIP [22] and Diff-CLIP+ (an improved version of Diff-CLIP). Recon1, Recon2 and Recon3 refer to inversion results via Restyle [1], DDIM and diffusion autoencoders, respectively. GAN-IPL and Diff-IPL denote integrating IPL with NADA and Diff-CLIP+, respectively. Real images are from CelebA-HQ dataset [29] and translated into two styles of images, "Wall painting" and "Tolkien elf". The yellow boxes show the key observation areas.

bism), while IPL presents consistently higher diversity and solve the mode collapse issue well. Our advantages mainly come from the fact that the latent mapper preserves sufficient image-specific and target-domain friendly features from the source-domain images. The produced prompt vectors provide more precise and diversified adaptation directions for the target-domain generator adaptation.

Quantitative comparison. To quantify the performance improvement of IPL compared to NADA [8], IS, SCS, ID and SIFID are evaluated. As reported in Tab.1, for IS, IPL outperforms NADA on all 10 settings, indicating our method achieves better image quality and diversity. For SCS and ID, IPL outperforms NADA on most of the 10 settings except "Human $\rightarrow$ Ukiyo-e". It is mainly because that "Ukiyo-e" naturally favors humans with narrow eyes and pale skin, which encourages identity changes during training. For SIFID, we collect 3 reference images (R_1, R_2, and R_3) on the internet for each target domain. Tab.1 shows that IPL outperforms NADA in most cases, indicating our superiority in generating precise target-domain styles.

User studies. For each target domain, 32 images generated by NADA and our method are provided to human observers, together with their corresponding source images and textual labels of target domains. Human observers are required to choose better synthesized images which are semantically more consistent with the target domain labels and preserve the useful source-domain information better. We collect 1210 responses from 121 people using a survey platform. As reported in the last column of Tab.1, 80.5% of users prefer our approach to NADA on average.

4.3. Real-world image translation

This task first inverts a real-world image to the latent code by a pre-trained inversion model and then feeds it to the trained target-domain generator to get the translated target-domain image. For GAN-based generators, we compare our method (GAN-IPL) with NADA by connecting the inversion model Restyle [1]. For diffusion model generators, we compare our method (Diff-IPL) with Diff-CLIP [22] and Diff-CLIP+ which is an improved version of Diff-CLIP [22] by replacing the original DDIM [42] with a diffusion autoencoder [35]. For these diffusion models, a deterministic inversion process is naturally provided.

As shown in Fig.5, comparing the results of NADA and GAN-IPL, IPL's superiority of alleviating mode collapse over NADA can still be observed. Comparing the results of Recon1, Recon2 and Recon3, diffusion models (Recon2 and Recon3) consistently perform better identity preservation than Restyle (Recon1) for real image inversion, especially for some uncommon stuffs in a human face photo, e.g., the hats, hands and tattoos in Fig.5. However, this property is not well inherited in the target domain generators with a fixed adaptation direction (see the results of Diff-CLIP and Diff-CLIP+). Our proposed IPL could help preserve the details in source images better and present the target-domain styles correctly (see the results of Diff-IPL). Quantitative evaluation results of Diff-CLIP, Diff-CLIP+ and Diff-IPL can be seen in supplementary materials.
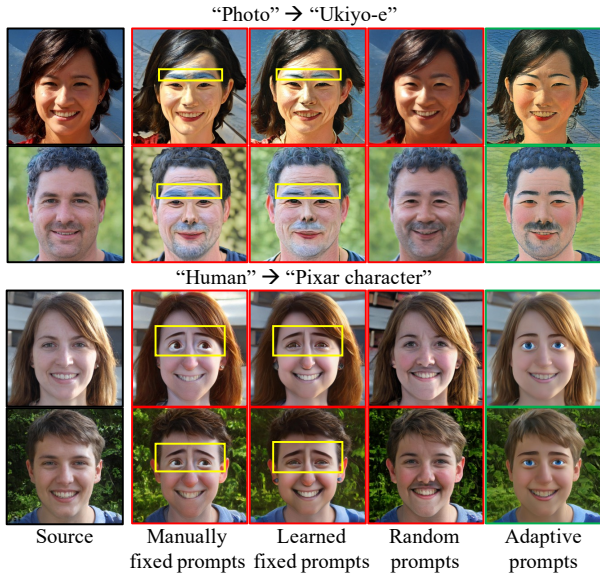
"Photo" → "Ukiyo-e"

"Human" → "Pixar character"

Source | Manually fixed prompts | Learned fixed prompts | Random prompts | Adaptive prompts

Figure 6. Ablation results of prompt designing schemes.

"Photo" → "Ukiyo-e"

"Human" → "Pixar character"

Source | $\lambda = 0$ | $\lambda = 1$ | $\lambda = 10$ | $\lambda = 20$

Figure 7. Ablation results of loss term ratios.

4.4. Ablation studies

Prompt designing schemes. We investigate four different prompt designing schemes: 1) manually fixed prompts (NADA), 2) learned fixed prompts, 3) random prompts and 4) adaptive prompts (Ours). Manually fixed prompts mean simply utilizing the manually designed prompts as NADA [8]. Learned fixed prompts denote unified prompt vectors produced by common prompt learning strategy [58] and shared for all images. Random prompts refer to prompt vectors produced by a randomly initialized latent mapper. Adaptive prompts denote the learned image-specific prompt vectors produced by our IPL method.

As illustrated in Fig.6, synthesized images with manually fixed prompts and learned fixed prompts show some similar mode collapse issues, e.g., blue eyebrows (Ukiyo-e) and depressed emotions (Pixar character). They both produce a fixed adaptation direction, which leads to identical supervision signals for all image pairs. Synthesized images with random prompts present more photo-realistic results but lack the desired target-domain style. A possible reason is that the random prompts contain some features conflicting with the target domain and impede the learning of the target domain style. Our adaptive prompts perform best since the prompts contain more image-specific and target-domain friendly features from the source-domain images.

Loss term ratios. We compare different values of the ratio parameter λ in Eq.(6), which is used to adjust the intensity of the domain regularization loss. Visual results are shown in Fig.7. In specific, when we set λ to a small value ($\lambda = 0$ as an extreme case), there is almost no constraint from the target domain. The learned prompts would excessively preserve the source-domain features. Thus the synthesized images are similar to their corresponding source
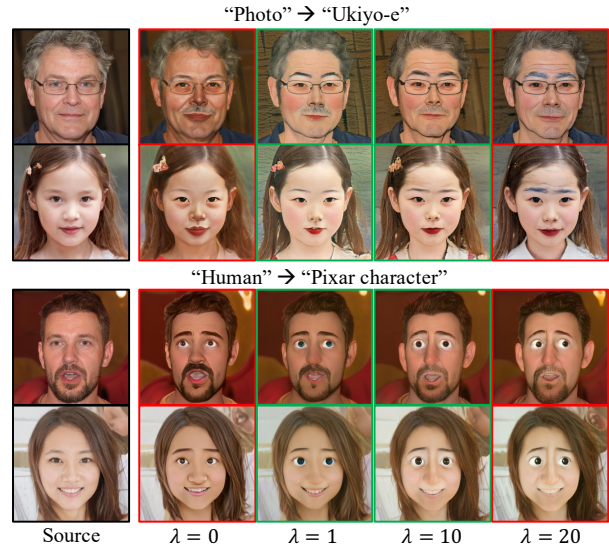
images. In contrast, if λ is set to a large value ($\lambda = 20$ as an example), a strong target-domain constraint will limit the diversity of the learned prompts. As a result, the synthesized images would slightly show some similar undesired patterns as images generated via fixed prompts. Therefore, in practical applications, λ should be a trade-off value (i.e., between 1 and 10).

5. Conclusion

In this paper, we have proposed a novel zero-shot generative model adaptation approach called **I**mage-specific **P**rompt **L**earning (IPL). In specific, we build a projection from latent codes to image-specific sets of prompt vectors via a latent mapper. With a contrastive learning scheme and a domain regularization constraint, the learned prompt vectors represent image-specific but target-domain-friendly features, producing more precise and diversified adaptation directions for target domain generator training. Compared with the state-of-the-art approaches, IPL consistently improves the quality of synthesized images and alleviates the mode collapse issue. Furthermore, IPL is independent of the type of generator and works well with both GANs and diffusion models, which exhibits good universality and adaptability. In the future, we will try to apply the proposed image-specific prompt learning strategy in other downstream tasks, such as unsupervised image captioning.

Acknowledgements

References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. ReStyle: A residual-based StyleGAN encoder via iterative refinement. In *ICCV*, 2021. 5, 7

[2] Sergey Bartunov and Dmitry Vetrov. Few-shot generative modelling with generative matching networks. In *AISTATS*, 2018. 3

[3] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 5, 6

[4] Louis Clouâtre and Marc Demers. FIGR: Few-shot image generation with reptile. *arXiv preprint arXiv:1901.02199*, 2019. 3

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 5, 6

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 5

[7] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 3

[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. In *SIGGRAPH*, 2022. 2, 3, 4, 5, 6, 7, 8

[9] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *ACL/IJCNLP*, 2021. 3

[10] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*, 2022. 3

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1

[12] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. TransFG: A transformer architecture for fine-grained recognition. In *AAAI*, 2022. 5, 6

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 3

[16] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *TACL*, 2020. 3

[17] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, 2021. 3

[18] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. 2, 3

[19] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 3

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 3

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 1, 2, 3, 5, 6

[22] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 2, 3, 5, 7

[23] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021. 3

[24] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 3

[25] Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing continuous prompts for generation. In *ACL/IJCNLP*, 2021. 3

[26] Weixin Liang, Zixuan Liu, and Can Liu. DAWSON: A domain adaptive few shot generation framework. *arXiv preprint arXiv:2001.00576*, 2020. 3

[27] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *WACV*, 2023. 2

[28] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. *arXiv preprint arXiv:2103.10385*, 2021. 3

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 7

[30] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning GANs. In *CVPR Workshops*, 2020. 2, 3

[31] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[32] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *ICCV*, 2019. 3

[33] 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 *CVPR*, 2021. 2, 3

[34] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian

Riedel. Language models as knowledge bases? In *EMNLP-IJCNLP*, 2019. 3

[35] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizad-wongsa, and Supasorn Suwajanakorn. Diffusion Autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022. 2, 5, 7

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 4, 5

[37] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *arXiv preprint arXiv:2010.11943*, 2020. 3

[38] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016. 5, 6

[39] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Sin-GAN: Learning a generative model from a single natural image. In *ICCV*, 2019. 5, 6

[40] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, 2020. 3

[41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 2, 3, 5, 7

[43] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 2019. 2

[44] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *TOG*, 2021. 5

[45] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for GAN training. *TIP*, 2021. 3

[46] Steven Walton, Ali Hassani, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. StyleNAT: Giving each head a new perspective. *arXiv preprint arXiv:2211.05770*, 2022. 2

[47] 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 *CVPR*, 2020. 2, 3

[48] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring GANs: generating images from limited data. In *ECCV*, 2018. 2, 3

[49] Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. Few shot generative model adaption via relaxed spatial structural alignment. In *CVPR*, 2022. 2, 3, 5, 6

[50] Xingqian Xu, Shant Navasardyan, Vahram Tadevosyan, Andranik Sargsyan, Yadong Mu, and Humphrey Shi. Image completion with heterogeneously filtered spectral hints. In *WACV*, 2023. 2

[51] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile Diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022. 2

[52] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *ICLR*, 2019. 3

[53] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. In *NeurIPS*, 2020. 2, 3

[54] Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for GANs. In *AAAI*, 2021. 3

[55] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for GAN training. *arXiv preprint arXiv:2006.02595*, 2020. 3

[56] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [MASK]: Learning vs. learning to recall. In *NAACL-HLT*, 2021. 3

[57] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 3

[58] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 3, 4, 8