Multi-Modal Face Stylization with a Generative Prior

Mengtian Li[1*] ⓘ , Yi Dong[2*] ⓘ , Minxuan Lin[1] ⓘ , Haibin Huang[1†] ⓘ , Pengfei Wan[1] ⓘ , and Chongyang Ma[1] ⓘ

[1]Kuaishou Technology, China   [2]Tsinghua University, China
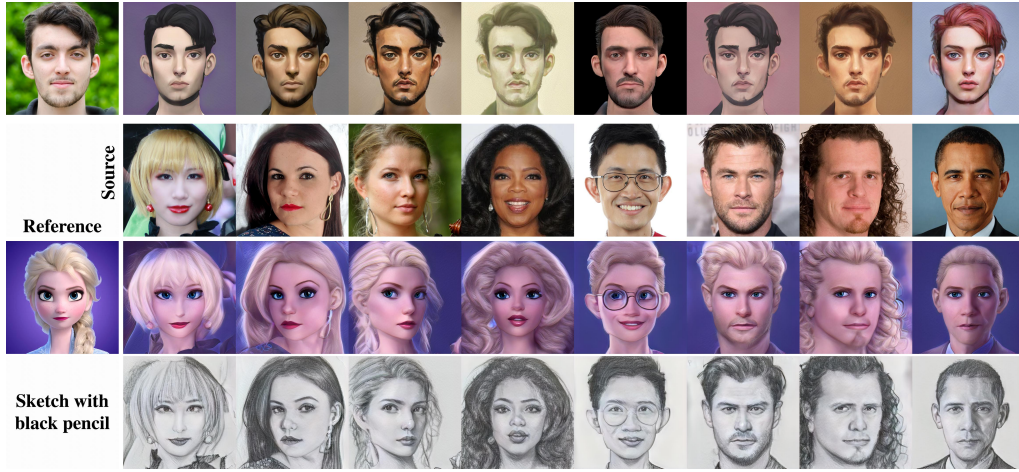
Figure 1: *Our proposed framework MMFS generates high-quality stylized faces with diverse styles (the top row) and can be applied to both one-shot and zero-shot stylization tasks (the third and fourth rows).*

Abstract
In this work, we introduce a new approach for face stylization. Despite existing methods achieving impressive results in this task, there is still room for improvement in generating high-quality artistic faces with diverse styles and accurate facial reconstruction. Our proposed framework, MMFS, supports multi-modal face stylization by leveraging the strengths of StyleGAN and integrates it into an encoder-decoder architecture. Specifically, we use the mid-resolution and high-resolution layers of StyleGAN as the decoder to generate high-quality faces, while aligning its low-resolution layer with the encoder to extract and preserve input facial details. We also introduce a two-stage training strategy, where we train the encoder in the first stage to align the feature maps with StyleGAN and enable a faithful reconstruction of input faces. In the second stage, the entire network is fine-tuned with artistic data for stylized face generation. To enable the fine-tuned model to be applied in zero-shot and one-shot stylization tasks, we train an additional mapping network from the large-scale Contrastive-Language-Image-Pre-training (CLIP) space to a latent w+ space of fine-tuned StyleGAN. Qualitative and quantitative experiments show that our framework achieves superior performance in both one-shot and zero-shot face stylization tasks, outperforming state-of-the-art methods by a large margin.

CCS Concepts
• *Computing methodologies* → *Image processing;*

1. Introduction

Artistic face stylization has been a popular research topic during the past few years, especially in the fields of computer graphics, computer vision, and machine learning [HYH*21, YNS19, YLLR19]. Given a facial image, artistic stylization methods aim to automatically transform it into a stylized version that is consistent with the input content but in a particular style. It can be used to create captivating visuals for various purposes, such as in the

---

*Joint first authors.
†Corresponding author: jackiehuanghaibin@gmail.com

entertainment industry, social media, and virtual reality [ALZ*23, Jam18, RBL*22].

However, achieving high-quality face stylization is a challenging task due to the complexity of facial structure and humans' perceptual sensitivity to subtle artifacts in the output. Compared to general image-to-image style transfer tasks, the prior knowledge humans have about facial features poses a significant challenge in preserving the original facial structure while applying the stylization process. Moreover, various downstream application scenarios can further complicate the stylization process, including situations where one model needs to support multiple styles, only a reference style image is given, or even just a simple textual description of the desired style is provided [CUYH20, KY22, LLYW18].

In this study, we aim to address the above challenges in generating diverse stylized faces and offer a unified framework called Multi-Modal Face Stylization (MMFS). MMFS allows for the generation of diverse stylized faces and provides control over the stylization process via either a single reference image or textual description. Our approach is based on three key observations. First, with the advancements in generative models, it is now possible to generate realistic and diverse human faces with fine-grained details using methods such as StyleGAN [KLA19, KLA*20]. We can leverage this generative prior as a strong basis for faithful face reconstruction during stylization. Second, we can improve this generative prior by fine-tuning it with large artistic face datasets such as AAHQ [LLQ*21], which enables us to support various styles using a single model. The architecture of StyleGAN2 also enables us to disentangle the latent space with respect to semantic attributes, enabling us to control the face content and style separately. Finally, we can leverage the semantic power of CLIP models, which allows us to align textual descriptions with the style latent space and enables us to support cross-domain guided stylization.

In light of these observations, we propose an encoder-decoder based architecture with a two-stage training strategy for our MMFS framework. Specifically, the core of MMFS is a StyleGAN-like generator, which is trained to accurately reconstruct realistic input faces in the first stage. Unlike previous methods that use w or $w+$ space for face reconstruction [AQW19, KLA*20], we propose to align the low-resolution layer of StyleGAN2 with a convolution based encoder. Compared with direct projection of the input image into a low-dimensional latent space, a convolution-based encoder preserves fine details in input facial images and results in better reconstructions. In practice, we use a pre-trained StyleGAN2 as our generative prior and fix its weights, while training the encoder to predict the layer at the 32×32 resolution in a self-reconstruction manner. Such architecture allows us to leverage the fine-grained details of StyleGAN2 while also preserving important features of the input image. In the second stage, we further fine-tune the entire network with artistic data for stylized face generation. The encoder is trained for style-free semantic and structural feature extraction and the decoder is trained to generate a stylized face from these structural features together with a latent style vector sampled from the latent w space of StyleGAN2. Our design also naturally enables the disentanglement of content and style which can be controlled by the encoder and the decoder separately.

After the two-stage training, we obtain a face stylization frame-

work that can transform a real face image into a randomly stylized version. To make the framework controllable for downstream tasks, we train an additional network to bridge the input image or textual guidance to the latent style vector. Specifically, we leverage the pre-trained CLIP model and adopt a four-layer transformer to map its space and the latent style space. Similar to the first stage, this mapping network can be trained using a self-reconstruction strategy, where random samples are generated for CLIP-w paired learning. Importantly, since the image and text are aligned in the CLIP feature space, the learned mapping network can be trained with images only but support textual guidance as well.

We conduct experiments to evaluate the effectiveness of MMFS in both one-shot and zero-shot stylization tasks. Our results demonstrate that MMFS is capable of generating high-quality stylized faces while preserving fine-grained details, as shown in Figure 1.

To summarize, our contributions are as follows:

- We propose MMFS, a novel framework which utilizes StyleGAN2 as a generative prior and integrate it into an encoder-decoder architecture. A two-stage training strategy is further designed to train MMFS for high-quality stylized face generation with diverse style support.
- MMFS is a flexible framework that can be adapted to various downstream tasks, including one-shot and zero-shot stylization. We also introduce a novel loss for better style preservation in one-shot stylizaton.
- Our experimental results demonstrate that MMFS generates better output than existing methods qualitatively and achieves state-of-the-art performance quantitatively.

2. Related Work

Face generative model. Recent StyleGAN methods [KLA19, KLA*20] have substantially improved face generation benchmarks. To accurately reconstruct and manipulate images of real faces, optimized-based GAN inversion methods [AQW19, AQW20] propose several constraint functions to obtain the latent directly. For example, II2S [ZAQ*20] introduces a P_N space to find the trade-off point of diversity and quality. However, inversion based methods are typically time-consuming. As a result, a series of encoder-based approaches are widely used to achieve fidelity and editability of synthetic portraits efficiently. pSp [RAP*21] and e4e [TAN*21] adopt a well-designed pretrained encoder network to map portrait images into the latent space. ReStyle [APCO21] utilizes iterative refinement to gradually modify the latent space vector via the residual-based encoder.

These inversion techniques have also been introduced for artistic face generation. Toonify [PA20] uses a layer swapping scheme to interpolate between real photos and cartoon images, which preserves the texture characteristic of the style and the structural fidelity of the photo. StyleCariGAN [JJJ*21] achieves caricature generation using shape exaggeration blocks to modulate StyleGAN2 feature maps. AniGAN [LZW*21] designs special fusion blocks and a double-branch discriminator to learn domain-specific and domain-shared information. AgileGAN [SLL*21] proposes a hierarchical VAE to encode the content into the $z+$ space to ensure consistency with the prior distribution. However, these methods only transmit spatially
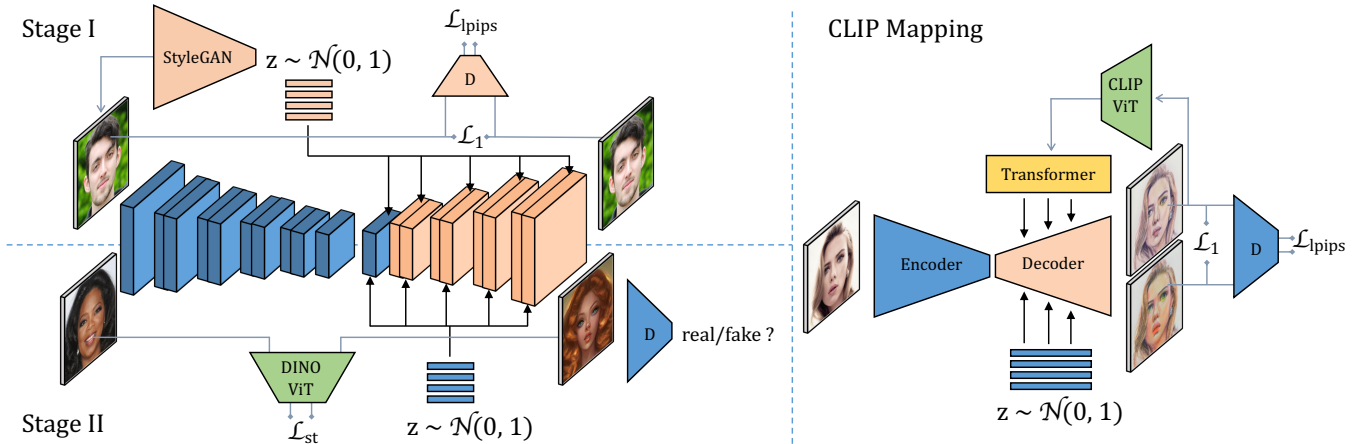
Figure 2: *Overview of our framework. Left: The proposed MMFS involves a two-stage training strategy, where Stage I trains an encoder to align with StyleGAN2 and to preserve fine-grained facial details, and Stage II fine-tunes the entire network for stylized face generation. Right: An additional mapping network is trained to bridge the CLIP feature space and the latent style space, providing a unified control for guided face stylization.*

structural signals through some stacks of one-dimensional vectors, which lead to generation results of limited fidelity. To address this issue, we propose a spatial encoder to modulate the feature map of StyleGAN2 directly.

Multi-modal face stylization. The diversity of generation results is a key feature of face stylization models. MUNIT [HLBK18] designs two encoder-decoder modules to exchange content and style features via adaptive instance normlization [HB17] to allow the style latent to be sampled. DRIT++ [LTM*20] enhances the disentanglement of representations by introducing a domain discriminator. StarGAN v2 [CUYH20] allows image and latent guided generation in a style branch to enrich the diversity of control condition. However, the decoder of these methods cannot achieve high-quality face stylization due to the lack of generative priors. More recently, BlendGAN [LLQ*21] transfers the scheme in StyleGAN2 decoder and achieves desirable face stylization results in terms of both quality and variety. Unfortunately, the decoupling ability is naturally limited by the $w/w+$ space of StyleGAN2, lacking of sufficient spatial information. To address these weaknesses, our approach explicitly decouples the content and style information in StyleGAN2, so that the content structure is not significantly affected while changing the style.

Guided face stylization. Given out-of-domain style reference images, the model should be able to adapt to the new domain. Some few-shot methods [LZLS20, MCS20, RCKH20, WGGB*20, WWH*18] focus on transferring to the target domain based on several reference images. For instance, Transferring GANs [WWH*18] defines the adaption of generation model by limited data as a transfer task. FreezeD [MCS20] freezes the lower layers of the discriminator to avoid overfitting. FSGA [OLL*21] proposes the CDC loss to keep the diversity of generated output. CtlGAN [WYT*22] constructs a novel encoder and the CDT loss to perform the stylization task.

Under a more strict setting of only one reference image, Jo-JoGAN [CF22] perturbs the part of inversion latent to augment

style. Mind-The-Gap [ZAFW22] constructs a geometric constraint relation in the CLIP space to shift domains while maintaining the overall structure. Similarly, OneshotCLIP [KY23] makes full use of the semantic consistency of the same portrait in the CLIP space via contrastive learning. DiFa [ZWJ*22] designs local and global adaptation functions to enhance the diversity and fidelity. Instead of fine-tuning the entire generator, GenDA [YSZ*21] imports two light modules called attribute adaptor and attribute classifier to refine the model. Generalized one-shot adaptation [ZLH*22] focuses on preserving decorations in a new domain. HyperNST [RGM*23] defines one-shot face stylization as a style transfer problem.

For zero-shot face stylization, the desired style information is described by input text, rather than a reference image. StyleGAN-NADA [GPM*22] uses the text direction in the CLIP space as guidance to optimize styles. StyleCLIP [PWS*21] adopts a latent mapper to hierarchically adjust the latent. TargetCLIP [CBPW22] adds an essence vector to the source latent to make multiple views in the CLIP space. With another usage of CLIP, CSLA [ZLG*22] projects the CLIP embedding into the latent space directly to edit image attributes. To provide feasibility under different conditions, our framework is designed to support both one-shot and zero-shot face stylization simultaneously.

3. Method

In this section, we formally introduce MMFS. As shown in Figure 2, the core of the network architecture is an encoder-decoder framework integrated with StyleGAN2, which has been pre-trained with FFHQ and serves as a generative prior for high-quality face generation. To ensure stable and efficient training of this network, we propose a two-stage training strategy. First, we train the encoder to approximate the low-resolution feature maps of the pre-trained StyleGAN2. Next, we fine-tune the entire encoder-decoder network with artistic portrait images, allowing the network to learn a domain translation from real faces to stylized ones. To enable guided face
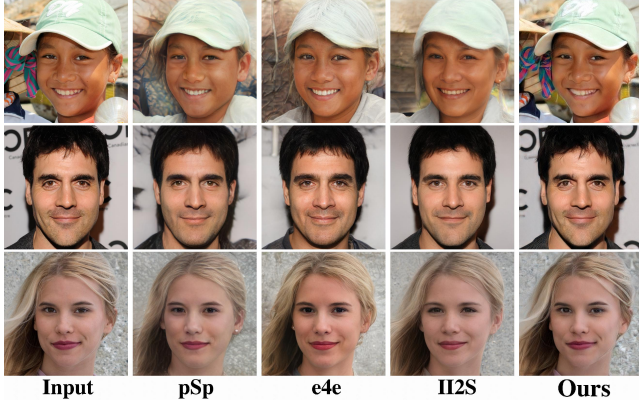
| **Input** | **pSp** | **e4e** | **II2S** | **Ours** |

Figure 3: *Comparison of reconstruction results. Our method can faithfully recover fine-grained details in local regions.*

stylization, we further train an additional mapping network from the CLIP feature space to the trained latent $w+$ space. This additional network provides a unified control for both one-shot and zero-shot face stylization.

3.1. Stage I: Encoder Pre-training

As demonstrated in [KLA19, KKC21], the mid- and high-resolution layers of StyleGANs are more related to semantic attributes and details of generated faces, while the low-resolution layers are responsible for controlling facial structure. Based on this observation, the first stage of MMFS is to train the encoder component of the network and to align its output with the 32×32 feature maps of StyleGAN2 to ensure faithful reconstruction of the input face.

During the training process, we sample a batch of images $\mathbf{I}_{ref}$ using the pre-trained StyleGAN2 generator $\mathcal{G}_{sty}$ with random noises z. These images and the corresponding noises are then fed into the encoder-decoder framework (**E** and **D**), resulting in reconstructed images $\mathbf{I}_{rec} = \mathbf{D}(\mathbf{E}(\mathcal{G}_{sty}(z)), z)$.

We use an $\mathcal{L}_1$ loss and a perceptual loss to penalize the difference between the reconstructed images and sampled images. To further preserve fine-grained details, we use the discriminator $\mathcal{D}$ of StyleGAN2 to evaluate the perceptual loss, as suggested by JoJoGAN [CF22]. Thus, the objective to optimize in this stage is:

$$\mathcal{L}_{\text{stage}-1} = \mathcal{L}_1(\mathbf{I}_{rec}, \mathbf{I}_{ref}) + \lambda_{perc} \sum_{l \in \{l_s\}} \mathcal{L}_1(\mathcal{D}^l(\mathbf{I}_{rec}), \mathcal{D}^l(\mathbf{I}_{ref})) \quad (1)$$

where λ_{perc} is the weight of the perceptual loss term and is set to 4.0 empirically, $\mathcal{D}^l(\cdot)$ denotes the l-th layer features extracted by the discriminator, and $\{l_s\}$ denotes the set of layers of the discriminator to compute the perceptual loss.

We compare the reconstruction results of our encoder-decoder network in Stage I with other GAN-inversion and StyleGAN2 encoding methods, including pSp [RAP*21], e4e [TAN*21], and II2S [ZAQ*20]. As shown in Figure 3, our method can faithfully recover fine-grained details including facial features and hair regions.

3.2. Stage II: Encoder and Decoder Fine-tuning

In the second training stage of MMFS, we utilize the self-reconstruction model obtained in the first stage and fine-tune the entire encoder-decoder network for face stylization. In this stage, our goal is to optimize the encoder to extract structural features that are independent of the desired artistic style, while making the decoder produce diverse stylized faces based on these structural features.

Towards this end, we leverage the following two key components. First, our decoder is StyleGAN-like, which can map a noise $z \in \mathcal{N}(0,1)$ into the latent space w and modulate convolution layers to produce diverse and high-fidelity results. Here z only controls the style, since the structural representation is obtained by the encoder from the input image. To generate a stylized face image from a real face image $\mathbf{I}_r$ with a noise z sampled from $\mathcal{N}(0,1)$, we use an adversarial loss to encourage our generated image to align with the distribution of stylized images $\mathbf{I}_s$:

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{I}_s}(\log \mathcal{D}(\mathbf{I}_s)) + \mathbb{E}_{\mathbf{I}_r,z}(\log(1 - \mathcal{D}(\mathbf{D}(\mathbf{E}(\mathbf{I}_r), z)))) \quad (2)$$

where $\mathcal{D}(\cdot)$ denotes the prediction of discriminator.

To further constrain the structure of the generated image to be similar to the input image, we use DINO-ViT [CTM*21] and its self-similarity metric. Specifically, we measure the self-similarity of DINO-ViT features to obtain the structure representation [TBTBD22]:

$$S^l(\mathbf{I})_{ij} = sim(K_i^l(\mathbf{I}), K_j^l(\mathbf{I})) \quad (3)$$

where $sim(\cdot, \cdot)$ is cosine similarity, and $K_i^l(\cdot)$ denotes the keys of the i-th token in the l-th layer of DINO-ViT. We measure the self-similarity for both generated image and input image, and encourage them to match each other, using the following structure loss:

$$\mathcal{L}_{st} = \|S^l(\mathbf{D}(\mathbf{E}(\mathbf{I}_r), z)) - S^l(\mathbf{I}_\mathbf{r})\|_F, \quad (4)$$

where $\| \cdot \|_F$ denotes the Frobenius norm. Here we adopt the last transformer layer of DINO-ViT.

The full objective in this stage for encoder and decoder fine-tuning is:

$$\mathcal{L}_{\text{stage}-2} = \mathcal{L}_{adv} + \lambda_{st} \mathcal{L}_{st} \quad (5)$$

where λ_{st} balances the two terms and is set to 0.5. We also leverage $\mathcal{R}_1$ regularization as in StyleGAN2 [KLA*20].

3.3. Guided Face Stylization

After the two-stage training, we obtain a fully trained encoder-decoder network that can randomly transform a realistic face image to a stylized version. However, this stylization process is random and beyond control. To enable more precise and controllable stylization, we propose to train an additional mapping network that can yield a latent code in the $w+$ space from a given style image.

Learning CLIP mapping. To enable guided stylization, we leverage a pre-trained CLIP model [RKH*21] to extract image features and train a mapping network to convert CLIP features to $w+$ latent codes. Due to the alignment of images and text in the CLIP feature

space, we can achieve text-guided stylization even if we only train on image data.

In our implementation, we use a four-layer transformer as the mapping network, with the input being the CLIP feature and a learnable positional embedding $\mathbf{e}_{\text{pos}} \in \mathbb{R}^{n_l \times d_c}$, where n_l is the number of latent codes in the $w+$ space, and d_c is the dimension of the CLIP feature. This mapping network is trained in a self-reconstruction manner, with the same loss function as the first stage (Section 3.1). To obtain a reference image for training, we randomly stylize an input image $\mathbf{I}_r$ with noise z: $\mathbf{I}_{ref} = \mathbf{D}(\mathbf{E}(\mathbf{I}_r), z)$. We then extract the CLIP features $\mathbf{F}_{\text{CLIP}}(\cdot)$ of $\mathbf{I}_{ref}$, feed them into the mapping network $\mathbf{M}(\cdot)$ to obtain the corresponding latent codes, and reconstruct the image as

$$\mathbf{I}_{rec} = \mathbf{D}(\mathbf{E}(\mathbf{I}_r), \mathbf{M}(\mathbf{F}_{\text{CLIP}}(\mathbf{I}_{ref}) + \mathbf{e}_{\text{pos}})). \tag{6}$$

The weight of the perceptual loss is set to $\lambda_{perc} = 4.0$ in this stage. The discriminator trained in the second stage (Section 3.2) is used to measure the perceptual difference.

Zero-shot stylization. Our encoder-decoder based network, which is equipped with a CLIP mapping, supports both text and image guided stylization in a single forward inference. However, due to the limited amount of training data, the network may not always generate high-quality images that match the style prompt. To further improve the stylization results, we introduce a fine-tuning progress that allows us to generate stylized images that are specific to a particular style prompt. In this fine-tuning stage, we use the encoder-decoder network as a prior, while the CLIP mapping plays a role in providing an initial solution.

Given a text prompt p_{text}, we obtain the $w+$ latent code using CLIP mapping: $c_{w+} = \mathbf{M}(\mathbf{F}_{\text{CLIP}}(p_{\text{text}}) + \mathbf{e}_{\text{pos}})$, and stylize an input image $\mathbf{I}_r$ with this latent code: $\mathbf{I}_s = \mathbf{D}(\mathbf{E}(\mathbf{I}_r), c_{w+})$. To measure the consistency between $\mathbf{I}_s$ and p_{text}, we use the directional CLIP loss [PWS*21]. Meanwhile, the structure of $\mathbf{I}_s$ should be consistent with $\mathbf{I}_r$. Thus, we optimize the following objective:

$$\mathcal{L}_{0-\text{shot}}(p_{\text{text}}) = \|S^l(\mathbf{I}_s) - S^l(\mathbf{I}_r)\|_F + \\ \lambda_c(1 - sim(\mathbf{d}_{\text{CLIP}}(\mathbf{I_s}), \mathbf{d}_{\text{CLIP}}(p_{\text{text}}))) \tag{7}$$

where λ_c is the weight of cosine similarity, and $\mathbf{d}_{\text{CLIP}}(\cdot)$ denotes the direction from the text/image anchor to CLIP features $\mathbf{F}_{\text{CLIP}}(\cdot)$. We freeze all the parameters except those of the decoder during this fine-tuning process.

One-shot stylization. For an image prompt p_{image}, we perform a fine-tuning step similar to the aforementioned zero-shot stylization. To further improve the quality of stylized image, we propose a new loss term that makes the best use of p_{image}. This new loss is built on the tokens of CLIP ViT image encoder. Specifically, we construct a group of orthogonal basis $U \in \mathbb{R}^{d_t \times \min(d_t, n_t)}$ (d_t and n_t denote the dimension and number of tokens) of ViT's tokens using SVD, which constitutes a feature space that is relevant to p_{image}. If the stylized image $\mathbf{I}_s$ has a style similar to the image prompt, the corresponding ViT tokens should remain unchanged as much as possible after being projected onto these orthogonal bases and then projected back. To achieve this goal, we add the following loss term to the fine-tuning

Figure 4: *Qualitative comparison on random stylization. MMFS achieves both good stylization and faithful facial detail preserving. DualStyleGAN [YJLL22a] requires an input exemplar as the reference and we randomly select four images from the AAHQ dataset as exemplars for stylization.*

objective:

$$\mathcal{L}_{1-\text{shot}} = \mathcal{L}_{0-\text{shot}}(p_{\text{image}}) + \\ \lambda_{proj}\|UU^T T_{\text{CLIP}}^l(\mathbf{I}_s) - T_{\text{CLIP}}^l(\mathbf{I}_s)\|_1 \tag{8}$$

where λ_{proj} controls the impact of the projection loss and $T_{\text{CLIP}}^l(\cdot)$ denotes the l-th layer's tokens of the CLIP ViT image encoder. We use the 4th layer in all of our experiments.

4. Experiments

Our proposed method, MMFS, is capable of generating stylized face images with high-quality and large diversity. In this section, we first describe our evaluation setups and then compare MMFS with several state-of-the-art baseline methods both qualitatively and quantitatively to evaluate the effectiveness of our approach in various stylization tasks.

4.1. Experimental Setup and Implementation Details

Baseline methods. We compare our MMFS with several previous methods on random stylization, including MUNIT [HLBK18], DRIT++ [LTM*20], and BlendGAN [LLQ*21]. All of them are trained on FFHQ [KLA19] and AAHQ [LLQ*21] datasets. For guided stylization, we conduct qualitative comparisons with MTG [ZAFW22], DiFa [ZWJ*22], JoJoGAN [CF22] and BlendGAN [LLQ*21] on one-shot stylization, and compare our method with StyleGAN-NADA [PWS*21] and CLIPstyler [KY22] on zero-shot stylization.

Evaluation metrics. We evaluate our method quantitatively from three perspectives: quality, identity-preservation, and diversity. To measure the quality, we use Frechet Inception Distance (FID) [HRU*17] to calculate the difference between the stylized images and samples from AAHQ. A lower FID score indicates a
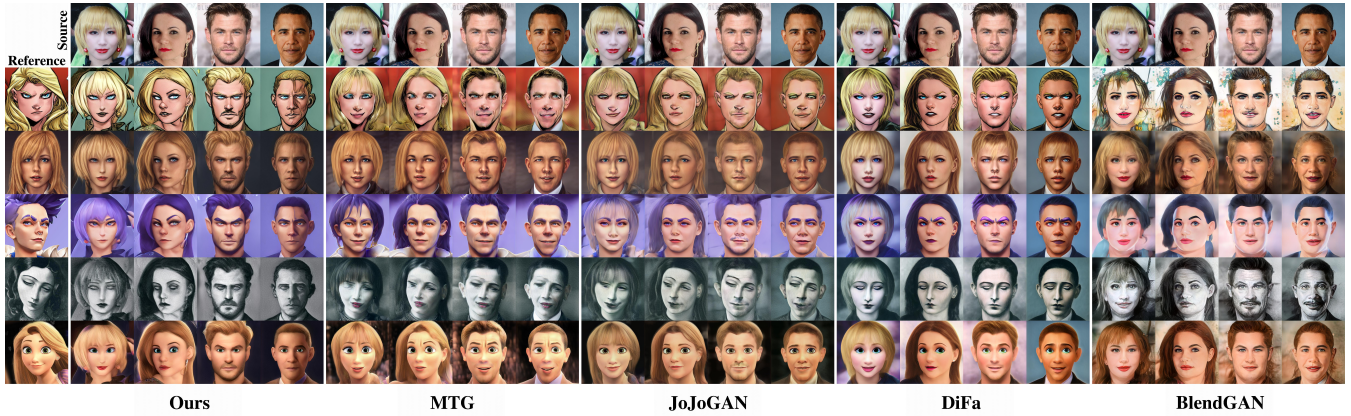
Figure 5: *Qualitative comparison results with several state-of-the-art methods on image-guided stylization.*

Figure 6: *Qualitative comparison results with several state-of-the-art methods on text-guided stylization. The source images are shown in the first row. The text prompts for the results from the second to the fifth rows are* 'a cubism style painting', 'pop art', 'watercolor painting', *and* 'painting in the style of Fernando Botero', *respectively.*

better stylization quality. For identity-preservation, we use the cosine similarity of Arcface [DGXZ19] features to evaluate the variation of face identity before and after stylization. To make this metric more intuitive, we use one minus the cosine similarity (Arcface-Dist). A lower Arcface-Dist value indicates better preservation of face identity. To measure diversity, we use LPIPS [ZIE*18] to calculate the perceptual difference between two randomly stylized images. A higher LPIPS value indicates larger diversity in stylization results.

Network training. We implement our method in PyTorch and train the network with Adam optimizer using a batch size of 8. We train the network with 10000, 90000, 60000 iterations at Stage I, Stage II, and CLIP mapping, respectively. For one-shot and zero-shot fine-tuning, we find 200 iterations are sufficient to achieve reasonable results and avoid over-fitting. In both Stages I and II, we set the learning rate to be 0.001, $\beta_1 = 0.1$ and $\beta_2 = 0.999$ for the optimizer. During the learning of CLIP mapping and zero/one-shot fine-tuning, the β_1 is set to 0.9 and the learning rate is decreased to 0.0002. Moreover, we apply an EMA decay of 0.999 (Stage I, Stage II, and CLIP mapping) or 0.99 (zero/one-shot fine-tuning) for our network except for the discriminator to stabilize the training process.

Timing statistics. The training process takes about 30 minutes per 1000 iterations. At test time, the image encoding, text encoding, random sampling, and one/zero-shot stylization step each takes 6.54, 6.59, 15.59, and 15.85 milliseconds on average, measured with single NVIDIA Tesla V100 GPU.

4.2. Random Stylization

We randomly select 10,000 face images from the FFHQ dataset, and feed them, in addition, the style code z sampled from $\mathcal{N}(0, 1)$, to each method. Table 1 shows the quantitative comparison between our approach and previous methods. Note that the blending indicator of BlendGAN is set to 6 for a trade-off between stylization strength and structural consistency. The stylized images of MUNIT and DRIT++ suffer from artifacts and blurriness while BlendGAN produces high-fidelity faces but of less noticeable style and lower diversity due to the above trade-off. In comparison, our method reaches a better balance of consistency and stylization strength, achieving both the best visual quality and diversity, as shown in Figure 4, thanks to the generative prior and our encoder-decoder based framework.

Table 1: *Quantitative comparison on random stylization. In each column, the best result is highlighted in **bold** while the second best number is marked with an underline. DualStyleGAN [YJLL22a] requires an input exemplar as the reference and we randomly sample 10000 images from the AAHQ dataset to obtain the corresponding numbers.*

Method	FID↓	Arcface-Dist↓	LPIPS↑
MUNIT [HLBK18]	18.96	0.511	0.505
DRIT++ [LTM*20]	<u>16.18</u>	0.514	0.489
BlendGAN [LLQ*21]	51.94	<u>0.509</u>	0.482
DualStyleGAN [YJLL22a]	54.96	0.582	**0.600**
MMFS	**10.39**	**0.507**	<u>0.583</u>

Figure 7: *Style interpolation results. In each row, from left to right, we present the input source image and interpolation results of two randomly sampled styles using the latent code of $w_1 \cdot \alpha + w_2 \cdot (1 - \alpha)$ where $\alpha \in \{1.0, 0.8, 0.6, 0.4, 0.2, 0.0\}$.*

4.3. Guided Stylization

One-shot stylization. In our evaluation of one-shot stylization, we compare our proposed MMFS with state-of-the-art methods, including MTG, JoJoGAN, DiFa, and BlendGAN. The results are shown in Figure 5. From the visual comparison, we can observe that MTG and JoJoGAN tend to generate twisty facial features when given slightly irregular style images, while the eye's orientation of their stylized faces is always under the influence of reference. The overall style strength of DiFa is relatively weaker than others except for BlendGAN, and the eyes rolling back frequently occurs in its stylized faces.

Additionally, all these methods suffer from a common dilemma: they are all built on StyleGAN2, and thus a GAN inversion is indispensable to project the input image onto the StyleGAN2 $w+$ space. Unfortunately, this inversion is not perfect and leads to the loss of structure. Furthermore, some methods (e.g., MTG and BlendGAN) rely on style-mixing, which can aggravate structural changes or weaken the style. In contrast, our proposed method
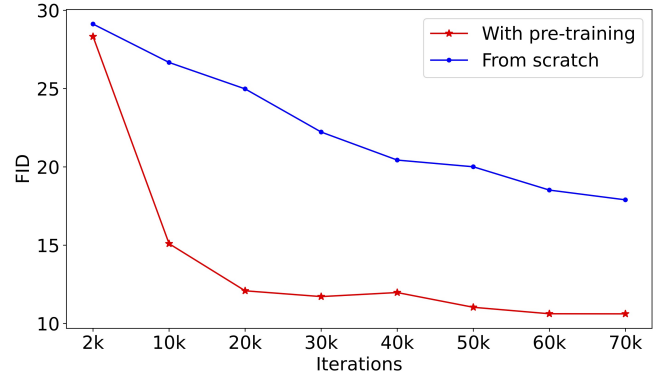
Figure 8: *FID curves of different training strategies. Our two-stage training leads to a significant boost in convergence speed and much lower FID.*
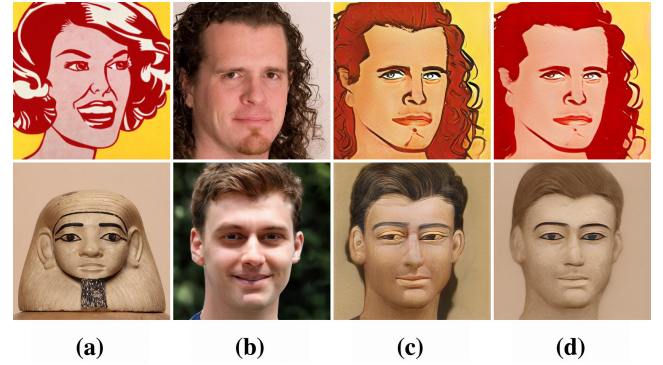

| (a) | (b) | (c) | (d) |

Figure 9: *Impact of fine-tuning and projection loss for one-shot stylization. From left to right: (a) reference styles; (b) input source images; (c) with fine-tuning but without the projection loss, and (d) our full method, respectively.*

integrates a StyleGAN2 into the encoder-decoder framework, allowing for both faithful reconstruction and high-quality stylization.

Zero-shot stylization. In Figure 6, we present qualitative comparisons between our proposed MMFS and state-of-the-art methods on zero-shot stylization. Similar to StyleGAN-based methods in one-shot stylization, StyleGAN-NADA suffers from significant modifications of structure and identity in its stylized faces. Furthermore, StyleGAN-NADA often generates unwanted colorful stripes in the background region of its generated images. Meanwhile, CLIPstyler adopts an encoder-decoder framework without the need for GAN inversion, ensuring more faithful reconstruction of input faces. However, its decoder and training strategy restrict its ability to generate high-quality face images, resulting in local artifacts. In contrast, our approach leverages a two-stage training strategy and integrates StyleGAN2 into the encoder-decoder framework, allowing for both faithful reconstruction of input faces and high-quality stylization, resulting in superior performance in zero-shot stylization. It is also worth mentioning that recent diffusion model based methods [DN21, ND21] have demonstrated superior performance for zero-shot stylization. However, it remains challenging

Mona Lisa painting

A sketch with crayon

(a) **(b)** **(c)** **(d)**

Figure 10: *Impact of fine-tuning for guided stylization. From left to right: (a) reference styles (top two rows) / input text (bottom two rows); (b) input source images; (c) results based on CLIP mapping without fine-tuning and (d) fine-tuning with directional CLIP loss, respectively.*

for these methods to maintain facial structure of input during the image-to-image translation process.

Style interpolation. Thanks to the StyleGAN-like decoder, our style code in the $w+$ space can be interpolated naturally. Figure 7 shows the visual results of style interpolation, in which the style of generated images gradually changes from one to another.

4.4. Ablation Study

Training strategy. To evaluate the advantage of our proposed training strategy, we conduct an experiment in which we remove the generative prior and train the entire encoder-decoder from scratch in Stage II. In Figure 8, we show the FID scores with respect to the training iterations. We observe that training from scratch leads to slower convergence and higher FID scores compared to our two-stage training strategy. By leveraging the pre-trained StyleGAN2 and the encoder in Stage I, our fine-tuning process in Stage II converges faster and achieves consistently lower FID scores.

Projection loss. Figure 9 demonstrates the positive impact of the projection loss proposed for one-shot stylization. While the directional CLIP loss encourages global semantic alignment between the stylized image and the style prompt, it may still not be sufficient enough to provide guidance for low-level features and fine-grained details. To further refine results in one-shot scenarios, our projection loss uses ViT tokens from the style reference to provide additional patch-level clues.

Figure 11: *Impact of different CLIP layers. From left to right: reference styles, input source images, results using the 4th, 8th, and 12th CLIP layer, respectively.*

Figure 12: *Our method has limitations in handling significant geometric deformations. Given a reference style such as a caricature (on the left), our method is unable to deform the facial structure accordingly (on the right) and instead transfers the source input with a similar texture appearance.*

Fine-tuning. As shown in Figure 10, the additional fine-tuning for zero-shot and one-shot stylization results in stylized images that better match the style prompt. When using only the CLIP mapping, the stylized results are reasonable, but the color tone and texture details may not always be consistent with the guidance. Fine-tuning allows for the direction of the text or image prompt in the CLIP space to provide a global cue to improve these issues.

CLIP features. We further conduct an experiment to validate the impact of different CLIP [RKH*21] layers in one-shot fine-tuning. As shown in Figure 11, shallower layers contain more low-level features, leading to better color match, while deeper layers are more semantic. We applied the proposed projection loss on deep layers which provides similar guidance as directional CLIP loss. We currently choose the 4th layer of CLIP features for a trade-off between low-level and semantic features, which works well in our experiments.

4.5. Limitations

Although MMFS achieves high-quality face stylization with large diversity in all the cases of random/one-shot/zero-shot stylization, there is still potential for further improvement. Specifically, our current implementation does not support significant geometric

Figure 13: *Out-of-distribution results. From left to right: input images, reconstruction results, and stylized output.*

deformation such as caricature, as shown in Figure 12. This issue may be addressed or alleviated by incorporating a learnable deformation module into the network. Moreover, our generated images are limited within the cropped region of FFHQ. Our method also has limited performance for input with large pose variations or significant occlusions that are out-of-distribution of FFHQ, as shown in Figure 13. Integrating our MMFS with VToonify [YJLL22b] is a promising future direction to unlock the power for full-size image stylization.

5. Conclusion

In this study, we present MMFS as a novel framework that leverages the strengths of StyleGAN2 and an encoder-decoder architecture to generate high-quality and diverse stylized faces while preserving fine-grained details. With our two-stage training strategy and CLIP-guided mapping network, MMFS offers a flexible and effective approach for one-shot and zero-shot stylization tasks. Our experimental results demonstrate that MMFS outperforms state-of-the-art methods by a large margin in both quantitative and qualitative evaluations. Overall, MMFS presents a promising solution for artistic face stylization with a wide range of potential applications. It is also worth exploring to expand this generative prior based framework to other applications beyond face stylization.

References

[ALZ*23] Abdal R., Lee H.-Y., Zhu P., Chai M., Siarohin A., Wonka P., Tulyakov S.: 3DAvatarGAN: Bridging Domains for Personalized Editable Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 4552–4562. 2

[APCO21] Alaluf Y., Patashnik O., Cohen-Or D.: ReStyle: A Residual-Based StyleGAN Encoder via Iterative Refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 6711–6720. 2

[AQW19] Abdal R., Qin Y., Wonka P.: Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 4432–4441. 2

[AQW20] Abdal R., Qin Y., Wonka P.: Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8296–8305. 2

[CBPW22] Chefer H., Benaim S., Paiss R., Wolf L.: Image-based clip-guided essence transfer. In *European Conference on Computer Vision* (2022), pp. 695–711. 3

[CF22] Chong M. J., Forsyth D.: Jojogan: One shot face stylization. In *European Conference on Computer Vision* (2022), pp. 128–152. 3, 4, 5

[CTM*21] Caron M., Touvron H., Misra I., Jégou H., Mairal J., Bojanowski P., Joulin A.: Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 9650–9660. 4

[CUYH20] Choi Y., Uh Y., Yoo J., Ha J.-W.: Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8188–8197. 2, 3

[DGXZ19] Deng J., Guo J., Xue N., Zafeiriou S.: Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 4690–4699. 6

[DN21] Dhariwal P., Nichol A.: Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems* (2021), pp. 8780–8794. 7

[GPM*22] Gal R., Patashnik O., Maron H., Bermano A. H., Chechik G., Cohen-Or D.: Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG) 41*, 4 (2022). 3

[HB17] Huang X., Belongie S.: Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2017), pp. 1501–1510. 3

[HLBK18] Huang X., Liu M.-Y., Belongie S., Kautz J.: Multi-modal unsupervised image-to-image translation. In *European Conference on Computer Vision* (2018), pp. 172–189. 3, 5, 7

[HRU*17] Heusel M., Ramsauer H., Unterthiner T., Nessler B., Hochreiter S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems* (2017). 5

[HYH*21] Han F., Ye S., He M., Chai M., Liao J.: Exemplar-based 3d portrait stylization. *IEEE Transactions on Visualization and Computer Graphics 29*, 2 (2021), 1371–1383. 1

[Jam18] Jamriska O.: Ebsynth: Fast example-based image synthesis and style transfer. https://github.com/jamriska/ebsynth, 2018. 2

[JJJ*21] Jang W., Ju G., Jung Y., Yang J., Tong X., Lee S.: Stylecarigan: caricature generation via stylegan feature map modulation. *ACM Transactions on Graphics (TOG) 40*, 4 (2021), 1–16. 2

[KKC21] Kang K., Kim S. H., Cho S.: Gan inversion for out-of-range images with geometric transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 13921–13929. 4

[KLA19] Karras T., Laine S., Aila T.: A style-based generator architecture for generative adversarial networks. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 4401–4410. 2, 4, 5

[KLA*20] KARRAS T., LAINE S., AITTALA M., HELLSTEN J., LEHTI-NEN J., AILA T.: Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8110–8119. 2, 4

[KY22] KWON G., YE J. C.: Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18062–18071. 2, 5

[KY23] KWON G., YE J. C.: One-shot adaptation of gan in just one clip. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023). 3

[LLQ*21] LIU M., LI Q., QIN Z., ZHANG G., WAN P., ZHENG W.: Blendgan: implicitly gan blending for arbitrary stylized face generation. In *Advances in Neural Information Processing Systems* (2021), pp. 29710–29722. 2, 3, 5, 7

[LLYW18] LIU A. H., LIU Y.-C., YEH Y.-Y., WANG Y.-C. F.: A unified feature disentangler for multi-domain image translation and manipulation. In *Advances in Neural Information Processing Systems* (2018). 2

[LTM*20] LEE H.-Y., TSENG H.-Y., MAO Q., HUANG J.-B., LU Y.-D., SINGH M., YANG M.-H.: Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision 128*, 10 (2020), 2402–2417. 3, 5, 7

[LZLS20] LI Y., ZHANG R., LU J., SHECHTMAN E.: Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780* (2020). 3

[LZW*21] LI B., ZHU Y., WANG Y., LIN C.-W., GHANEM B., SHEN L.: Anigan: Style-guided generative adversarial networks for unsupervised anime face generation. *IEEE Transactions on Multimedia 24* (2021), 4077–4091. 2

[MCS20] MO S., CHO M., SHIN J.: Freeze the discriminator: a simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964* (2020). 3

[ND21] NICHOL A. Q., DHARIWAL P.: Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)* (2021), pp. 8162–8171. 7

[OLL*21] OJHA U., LI Y., LU J., EFROS A. A., LEE Y. J., SHECHTMAN E., ZHANG R.: Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 10743–10752. 3

[PA20] PINKNEY J. N., ADLER D.: Resolution dependent gan interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334* (2020). 2

[PWS*21] PATASHNIK O., WU Z., SHECHTMAN E., COHEN-OR D., LISCHINSKI D.: Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 2085–2094. 3, 5

[RAP*21] RICHARDSON E., ALALUF Y., PATASHNIK O., NITZAN Y., AZAR Y., SHAPIRO S., COHEN-OR D.: Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 2287–2296. 2, 4

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10684–10695. 2

[RCKH20] ROBB E., CHU W.-S., KUMAR A., HUANG J.-B.: Few-shot adaptation of generative adversarial networks. *arXiv preprint arXiv:2010.11943* (2020). 3

[RGM*23] RUTA D., GILBERT A., MOTIIAN S., FAIETA B., LIN Z., COLLOMOSSE J.: HyperNST: Hyper-Networks for Neural Style Transfer. In *European Conference on Computer Vision* (2023), pp. 201–217. 3

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J.,

ET AL.: Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)* (2021), PMLR, pp. 8748–8763. 4, 8

[SLL*21] SONG G., LUO L., LIU J., MA W.-C., LAI C., ZHENG C., CHAM T.-J.: Agilegan: stylizing portraits by inversion-consistent transfer learning. *ACM Transactions on Graphics (TOG) 40*, 4 (2021), 1–13. 2

[TAN*21] TOV O., ALALUF Y., NITZAN Y., PATASHNIK O., COHEN-OR D.: Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG) 40*, 4 (2021), 1–14. 2, 4

[TBTBD22] TUMANYAN N., BAR-TAL O., BAGON S., DEKEL T.: Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10748–10757. 4

[WGGB*20] WANG Y., GONZALEZ-GARCIA A., BERGA D., HERRANZ L., KHAN F. S., WEIJER J. V. D.: 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* (2020), pp. 9332–9341. 3

[WWH*18] WANG Y., WU C., HERRANZ L., VAN DE WEIJER J., GONZALEZ-GARCIA A., RADUCANU B.: Transferring gans: generating images from limited data. In *European Conference on Computer Vision* (2018), pp. 218–234. 3

[WYT*22] WANG Y., YI R., TAI Y., WANG C., MA L.: CtlGAN: Few-shot Artistic Portraits Generation with Contrastive Transfer Learning. *arXiv preprint arXiv:2203.08612* (2022). 3

[YJLL22a] YANG S., JIANG L., LIU Z., LOY C. C.: Pastiche Master: Exemplar-Based High-Resolution Portrait Style Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 7693–7702. 5, 7

[YJLL22b] YANG S., JIANG L., LIU Z., LOY C. C.: Vtoonify: Control-lable high-resolution portrait video style transfer. *ACM Transactions on Graphics (TOG) 41*, 6 (2022), 1–15. 9

[YLLR19] YI R., LIU Y.-J., LAI Y.-K., ROSIN P. L.: Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 10743–10752. 1

[YNS19] YANIV J., NEWMAN Y., SHAMIR A.: The face of art: landmark detection and geometric style in portraits. *ACM Transactions on Graphics (TOG) 38*, 4 (2019), 1–15. 1

[YSZ*21] YANG C., SHEN Y., ZHANG Z., XU Y., ZHU J., WU Z., ZHOU B.: One-shot generative domain adaptation. *arXiv preprint arXiv:2111.09876* (2021). 3

[ZAFW22] ZHU P., ABDAL R., FEMIANI J., WONKA P.: Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. In *International Conference on Learning Representations* (2022). 3, 5

[ZAQ*20] ZHU P., ABDAL R., QIN Y., FEMIANI J., WONKA P.: Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036* (2020). 2, 4

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018). 6

[ZLG*22] ZHENG W., LI Q., GUO X., WAN P., WANG Z.: Bridging clip and stylegan through latent alignment for image editing. *arXiv preprint arXiv:2210.04506* (2022). 3

[ZLH*22] ZHANG Z., LIU Y., HAN C., GUO T., YAO T., MEI T.: Generalized one-shot domain adaptation of generative adversarial networks. In *Advances in Neural Information Processing Systems* (2022), pp. 13718–13730. 3

[ZWJ*22] ZHANG Y., WEI Y., JI Z., BAI J., ZUO W., ET AL.: Towards diverse and faithful one-shot adaption of generative adversarial networks. In *Advances in Neural Information Processing Systems* (2022), pp. 37297–37308. 3, 5