Domain Generalizable Portrait Style Transfer

Xinbo Wang[1†]   Wenju Xu[2†]   Qing Zhang[1,3*]   Wei-Shi Zheng[1,3]

[1] School of Computer Science and Engineering, Sun Yat-sen University, China   [2] AMAZON
[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China

wangxb29@mail2.sysu.edu.cn   xuwenju123@gmail.com
zhangq93@mail.sysu.edu.cn   wszheng@ieee.org

Figure 1. **Domain generalizable portrait style transfer.** By training on the CelebAMask-HQ dataset with only 30K portrait photos, our method allows for high-quality semantic-aware style transfer between any two portraits from a wide variety of domains including photo, cartoon, sketch, animation, etc.

Abstract

This paper presents a portrait style transfer method that generalizes well to various different domains while enabling high-quality semantic-aligned stylization on regions including hair, eyes, eyelashes, skins, lips, and background. To this end, we propose to establish dense semantic correspondence between the given input and reference portraits based on a pre-trained model and a semantic adapter, with which we obtain a warped reference semantically aligned with the input. To ensure effective yet controllable style transfer, we devise an AdaIN-Wavelet transform to balance content preservation and stylization by blending low-frequency information of the warped reference with high-frequency information of the input in the latent space. A style adapter is also designed to provide style guidance from the warped reference. With the stylized latent from AdaIN-Wavelet transform, we employ a dual-conditional diffusion model that integrates a ControlNet recording high-frequency information and the style guidance to generate the final result. Extensive experiments demonstrate the superiority of our method. Our code and trained model are available at https://github.com/wangxb29/DGPST.

1. Introduction

Portrait style transfer has gained popularity as an advanced image editing technique, providing an automated way to apply the visual style of a reference portrait to an input portrait. This task is particularly challenging as it requires precise local tone adjustments across different facial regions, such as the skin, lips, eyes, hair, and background, to faithfully replicate the reference style while preserving the portrait identity and facial structure details. As illustrated in Figure 1, this technique is expected to enable a wide range of transformations, such as converting a human portrait into various artistic styles, modernizing old photographs, and adding color to sketch images, allowing casual photographers to effortlessly transform their photos into stunning

[†]Equal contribution.
[*]Corresponding author.

masterpieces inspired by their favorite artworks.

To enhance the efficiency of portrait style transfer, several prior works [1, 7, 28, 29, 32, 34, 38, 41, 48] have demonstrated compelling results. However, these methods are not easily adaptable for generating portraits across multiple domains. More recently, diffusion models have been leveraged for various applications, including semantic matching [15, 33, 43] and style transfer [2, 3, 5, 10, 35, 36, 39, 47]. While techniques such as those proposed in [1, 16, 28, 29] aim to achieve stylization by aligning semantic regions, they are effective only when the structural differences between the input and reference portraits are minimal. When applied to portraits with significant structural variations, these methods struggle to transfer style effectively between corresponding semantic regions.

In this paper, we propose a novel domain-generalizable diffusion model for portrait style transfer. Our approach first warps the reference portrait using correspondences learned from diffusion models, producing a high-quality warped style reference. To generate realistic and visually coherent portrait, our framework is designed to integrate a dual-conditional diffusion model, leveraging ControlNet to extract high-frequency details from the input image for structural guidance, while an image adapter utilizes the warped reference to provide style guidance.

Unlike previous diffusion-based methods [2, 3, 5, 36, 47], our approach establishes dense semantic-aligned correspondences to enhance the quality of style transfer. Specifically, we introduce a semantic adapter, comprising a frozen diffusion U-Net and a CLIP encoder to extract image features from both the content and style reference images. As naively using these extracted features for correspondence estimation can result in incomplete semantic regions due to misalignment, we introduce mask warping loss and cyclic warping consistency loss for optimizing the semantic adapter to encourage semantic consistency in the extracted features. With this refined representation, our method enables accurate semantic-aware style transfer while ensuring the visual coherence in the generated portraits.

Color tone plays a crucial role in defining artistic style during portrait style transfer. If the generation process starts from the latent representation of the input image, the resulting portrait tends to retain its original color tone, limiting the effectiveness of style transfer. To better integrate structure and style, we propose a novel AdaIN-Wavelet Transform for latent initialization. Specifically, we first compute the latent representation of the warped reference image using DDIM inversion, as initializing from this latent naturally enhances color transfer effect. However, since the warped reference image may lose some fine-scale content details, directly using its latent representation may result in blurred portrait. To mitigate this, we blend the low-frequency information of the warped reference latent with

the high-frequency information of the input latent, achieving a balance between stylization and content preservation.

With the aforementioned designs—learning domain-agnostic semantic correspondences for portraits and the AdaIN-Wavelet latent initialization—our model enables effective portrait style transfer across various different domains. Through extensive experiments, we demonstrate the effectiveness of our approach in various artistic styles, modernizing old photographs, and colorizing sketch images. Our key contributions can be summarized as follows:

- We present a portrait style transfer framework that can robustly process portraits from various different domains by training on a small-scale real-world portrait dataset.

- We propose to utilize pre-trained diffusion model and also a semantic adapter for robust dense correspondence extraction. In addition, we develop a AdaIN-Wavelet transform to balance stylization and content preservation.

- Extensive experiments show that our method significantly outperforms previous methods in both visual quality and quantitative metrics.

2. Related Work

Diffusion-based Style Transfer. Diffusion models have been explored for style transfer, with a common approach leveraging cross-attention layers in text-to-image diffusion models. These methods can be broadly categorized into three groups: (i) directly using text prompts [37], (ii) mapping the reference image into textual embeddings [46, 47], and (iii) incorporating an additional cross-attention mechanism with image prompts [24, 35, 42]. Beyond these approaches, training-free methods have also been developed based on pre-trained text-to-image diffusion models. These methods inject features into self-attention layers [3, 5, 17] during the generation process to guide stylization. Additionally, StyleDiffusion [39] fine-tunes a diffusion model for a given reference image using a CLIP-based style disentanglement loss. However, most existing diffusion-based methods are designed for artistic style transfer that does not need to consider semantic correspondence, and thus basically struggle to deal with portrait style transfer involving strong semantic relevance, especially when the portraits are from different domains.

Portrait Style Transfer. Significant progress has recently been made in this field. Shih et al. [28] pioneered photorealistic portrait style transfer by first warping the reference portrait to align with the input portrait, followed by a multiscale local contrast transfer. Concurrently, Shu et al. [29] redefined portrait relighting as a mass transport problem and proposed an algorithm to address it. With the advent of CNN-based neural style transfer [6, 21], specialized methods for portrait style transfer emerged, including pho-
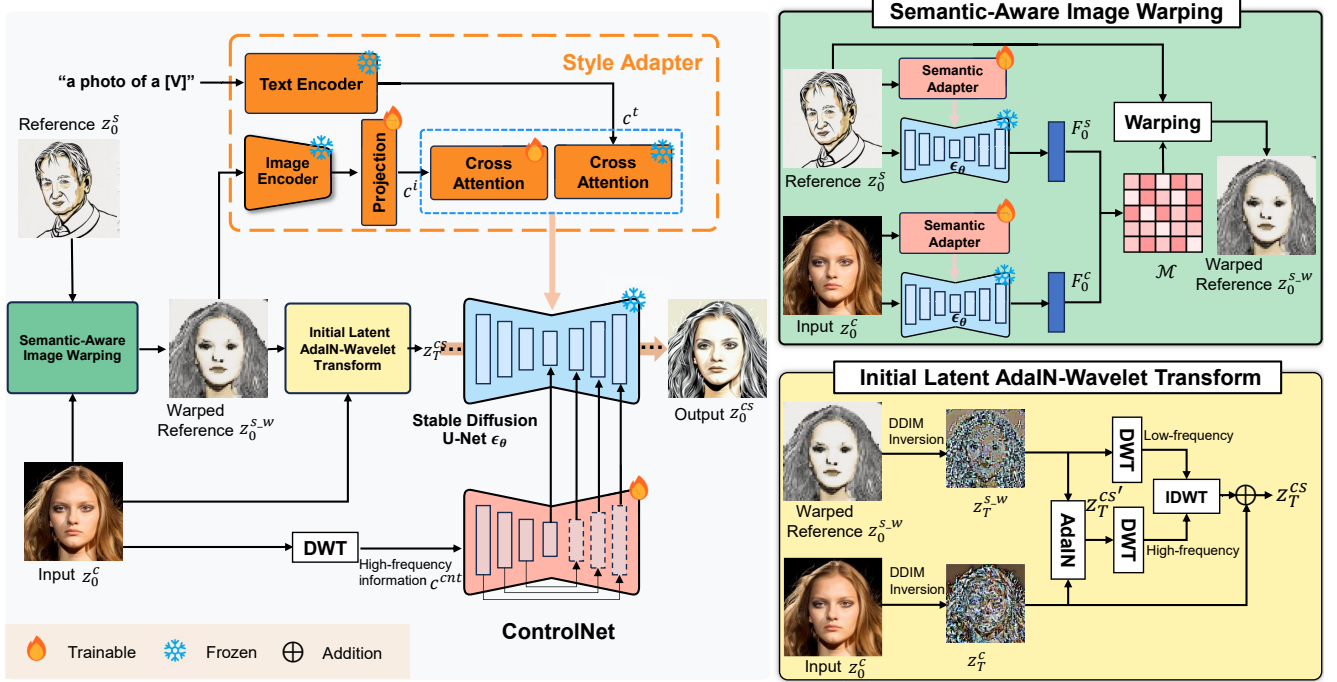
Figure 2. **Framework overview.** (Left) Illustration of the proposed portrait style transfer method. We first feed the input and reference portraits into Stable Diffusion and the Semantic Adapter. Next, we extract diffusion features and compute a correlation matrix, which is used to warp the reference portrait. During sampling, we provide structure guidance by extracting the high-frequency information of the input portrait using Discrete Wavelet Transform and feed it into ControlNet. For style guidance, the warped reference is input into the style adapter. To obtain the initial latent, we perform DDIM inversion on both the warped reference and input portraits, followed by applying AdaIN between them. The final initial latent is obtained by combining the low-frequency information of the warped reference latent with the high-frequency information from the input latent after AdaIN.

torealistic [8] and artistic styles [27]. Among them, Guo and Jiang [8] enhanced spatial transformations by maximizing the cross-correlation of features within segmented regions between the input and reference images. GAN-based methods have also gained popularity in face stylization due to their ability to generate high-quality, highly detailed images [1, 7, 34]. Notably, some approaches leverage semantic masks to enable style transfer for specific facial regions [7, 34]. Moreover, StyleGAN [11, 12], known for its impressive high-resolution face synthesis and hierarchical style control, has served as a foundation for several recent artistic portrait style transfer methods [30, 40, 41, 48]. However, these methods are unsuitable for our task, as they inevitably alter portrait identity.

3. Preliminary

Diffusion model is a class of generative models consisting of a T-step forward process and a T-step reverse process. Here, T represents the total number of steps in the forward process. At each timestep t, the forward process adds Gaussian noise to the input according to a predetermined noise schedule α_t:

$$z_t = \sqrt{\alpha_t} \cdot z_0 + \sqrt{1 - \alpha_t} \cdot \epsilon, \epsilon \sim \mathcal{N}(0, I). \quad (1)$$

A corresponding reverse process involves predicting the added noise ϵ using a network ϵ_θ, which takes z_t and timestep t as input, and then samples z_{t-1} from z_t using a formula, such as DDPM [9] and DDIM [31]. The ϵ_θ is usually parameterized as a U-Net for image generation. The added noise is progressively removed, ultimately resulting in a clean image. In this paper, we implement our method using Stable Diffusion (SD) [26], a large-scale pre-trained text-to-image model. The training objective of SD is defined as:

$$\mathcal{L} = \mathbb{E}_{z_t, t, y, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon - \epsilon_\theta(z_t, t, y) \|_2^2 \right], \quad (2)$$

where y is the embedding of text prompts generated by a pre-trained CLIP [25] text encoder.

4. Method

Our method aims to achieve semantically aligned portrait style transfer while preserving content details. Since our

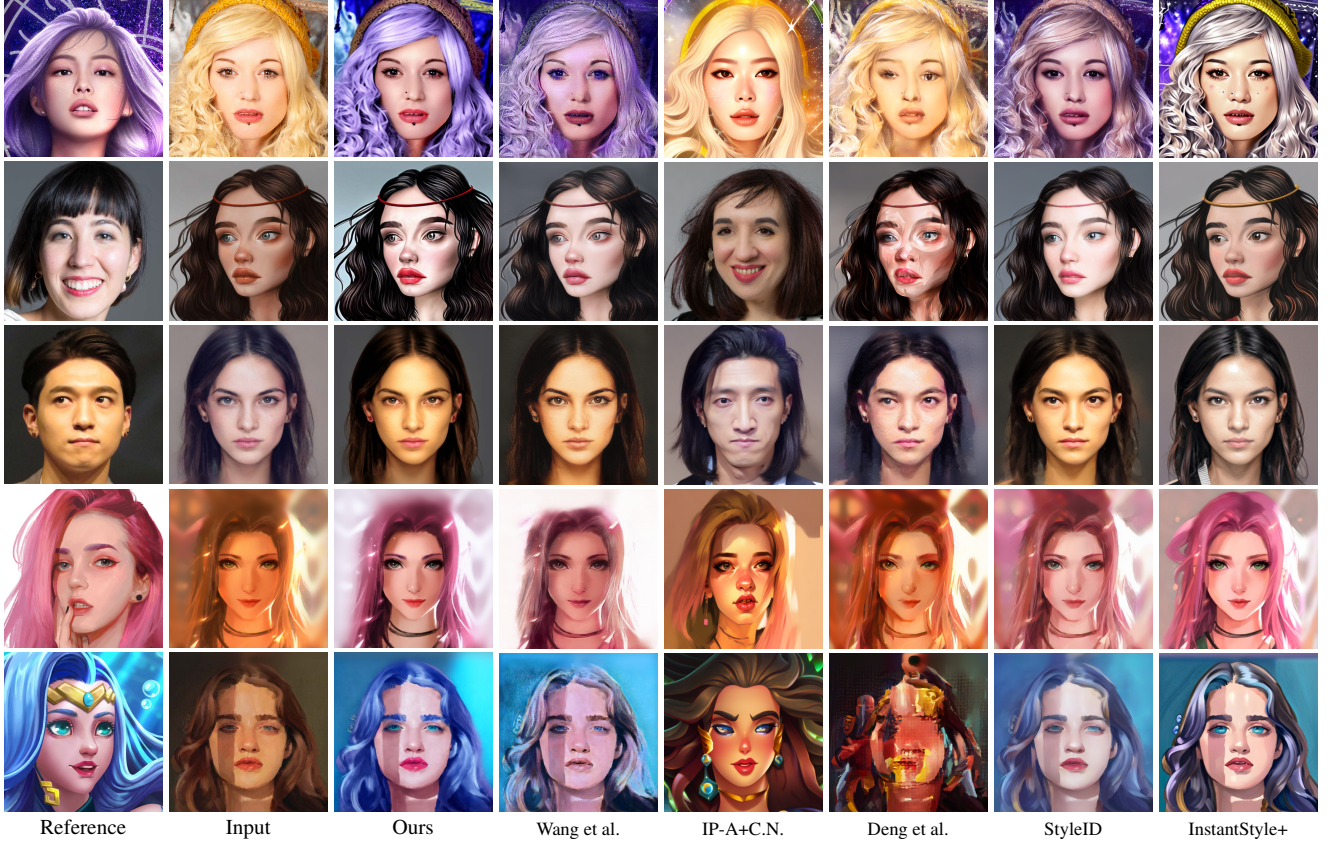| Reference | Input | Ours | Wang et al. | IP-A+C.N. | Deng et al. | StyleID | InstantStyle+ |

Figure 3. **Qualitative comparison with SOTA portrait style transfer methods.**

focus is on detailing the proposed method within the diffusion process, we denote the content image as z_0^c and the style image as z_0^s. As illustrated in Figure 2, our approach first warps the portrait style image using the correspondence learned from diffusion models to obtain a high-quality warped style reference $z_0^{s\text{-}w}$ (Section 4.1). To generate a realistic portrait, our framework incorporates a dual-conditional diffusion model (Section 4.2). This model utilizes the ControlNet ϵ_θ, which extracts high-frequency information from the content image z_0^c to provide content guidance, while an image adapter leverages the warped reference $z_0^{s\text{-}w}$ to offer style guidance. Additionally, we introduce a latent initialization strategy (Section 4.3) to achieve a better balance between stylization and content preservation.

4.1. Semantic-Aware Style Alignment

Different from previous diffusion-based methods [33, 43] for correspondence extraction, our approach focuses on establishing dense correspondence across portraits to achieve high-quality portrait style transfer. By ensuring style transfer between semantically aligned regions, we enhance both accuracy and visual coherence. This section details our method for constructing robust semantic correspondence to enable precise and effective style alignment.

Correspondence Extraction. We first utilize a semantic adapter and a pre-trained CLIP image encoder to extract image features. Specifically, we obtain features from the penultimate layer of the encoder, which consists of a total of $16 \times 16 + 1 = 256$ tokens with one class token and 256 tokens corresponding to individual image patches. These tokens are then processed through a projection network and fed into Stable Diffusion U-Net for decoupled cross-attention [42]. Next, we feed z_0^c and z_0^s into the Stable Diffusion U-Net and extract features from the third upsampling block of the U-Net, represented as $F_0^c \in \mathbb{R}^{HW \times C}$ and $F_0^s \in \mathbb{R}^{HW \times C}$ for images z_0^c and z_0^s (H and W represent the feature spatial sizes, while C is the number of channels), respectively. Then, the element of i_{th} row and j_{th} column in the correlation matrix is computed as:

$$\mathcal{M}(i,j) = \frac{(F_0^c(i) - \mu_{F_0^c(i)}) \cdot (F_0^s(j) - \mu_{F_0^s(j)})}{\|F_0^c(i) - \mu_{F_0^c(i)}\|_2 \, \|F_0^s(j) - \mu_{F_0^s(j)}\|_2}, \quad (3)$$

where $\mu_{F_0^c} \in \mathbb{R}^{HW}$ and $\mu_{F_0^s} \in \mathbb{R}^{HW}$ are mean vectors of F_0^c and F_0^s along the channel dimension.

Semantic-Aware Image Warping. With the semantic correspondence, we perform a differentiable warping of the reference portrait z_0^s toward the input z_0^c, resulting in $z_0^{s\text{-}w}$.

Each pixel of $z_0^{s\text{-}w}$ is computed as a weighted average of z_0^s:

$$z_0^{s\text{-}w}(i) = \sum_j \operatorname*{softmax}_j(\mathcal{M}(i,j)/\tau) \cdot z_0^s(j), \qquad (4)$$

where τ is set to 0.01 to make the softmax curve sharper. To further enhance the accuracy of semantic correspondence, we introduce a mask warping loss during training, defined as:

$$\mathcal{L}_{mask} = \|M^c - M^{s\text{-}w}\|_1, \qquad (5)$$

where M^c represents the semantic mask of z_0^c, and $M^{s\text{-}w}$ is obtained by warping the mask M^s of z_0^s following Equation 4. Besides, we introduce a cyclic warping consistency loss to reinforce style preservation during the warping process:

$$\mathcal{L}_{cwc} = \mathcal{L}_{LPIPS}(z_0^s, z^{s'\text{-}w}), \qquad (6)$$

where $z^{s'\text{-}w}$ is obtained by warping $z^{s\text{-}w}$ using the transposed correlation matrix $\mathcal{M}^\top$. By incorporating these two losses, the dense correspondence learned from diffusion models is further refined and enhanced. Figure 4 compares the effect of replacing the correspondence in our method with two alternatives [33]. As shown in the warped reference, output, and visualized correspondence map, the correspondence in our method is more accurate. This accuracy enables our method to generate portrait images that faithfully replicate the reference style while preserving the portrait's identity and facial structure details.

4.2. Dual Conditional Diffusion Model

The proposed dual-conditional diffusion model leverages both structure guidance and style guidance to generate realistic portrait images. This novel architectural design enables the integration of high-frequency structures and semantically aligned style references, enhancing the quality and coherence of the generated images.

Structure Guidance. Unlike the previous method [36], our model utilizes the high-frequency information from the content image z_0^c through ControlNet to provide style-independent structural guidance. To achieve this, we apply the Haar Discrete Wavelet Transform (DWT) to extract the high-frequency components of the content image. Specifically, the Haar wavelet transform performs a depth-wise convolution with four kernels, $\{LL^\top\ LH^\top\ HL^\top\ HH^\top\}$, where $L = \frac{1}{\sqrt{2}}[1\ \ 1]^\top$, $H = \frac{1}{\sqrt{2}}[-1\ \ 1]^\top$. Among these, $LL^\top$ acts as a low-pass filter, while $LH^\top$, $HL^\top$, and $HH^\top$ serve as high-pass filters that capture high-frequency information in the horizontal, vertical, and diagonal directions, respectively. We use the outputs from the three high-pass filters $LH^\top$, $HL^\top$ and $HH^\top$ as the corresponding input to ControlNet, which can be formulated as:

$$c^{cnt} = [z_{LH}^c, z_{HL}^c, z_{HH}^c] \qquad (7)$$
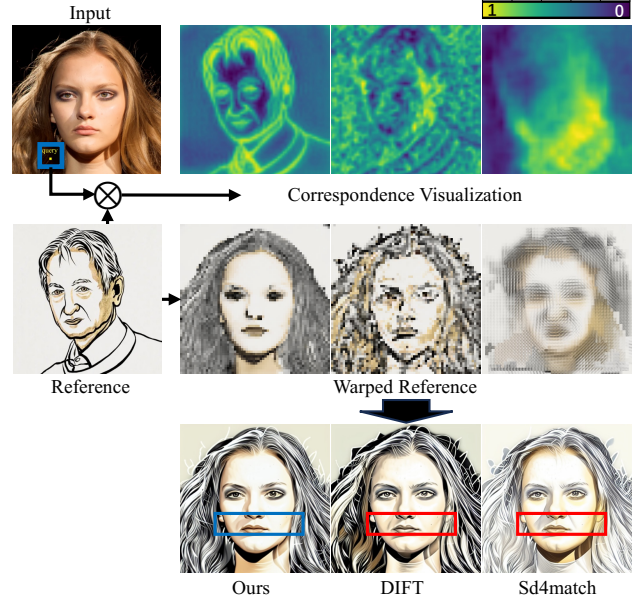$$= \text{Conv}([LH^\top, HL^\top, HH^\top], z_0^c). \qquad (8)$$

Figure 4. **Effect of our semantic correspondence.** For a query point (yellow point in input), the first row shows the similarity map computed by different methods. The second row displays the warped references. The third row shows the outputs generated using these different warped references. As seen, our correspondence achieves higher similarity in the semantic region of the reference, demonstrating its superior accuracy.

Note that the original spatial size of c^{cnt} is half that of z_0^c. Therefore, we upsample c^{cnt} using bilinear interpolation and use the upsampled c^{cnt} as input to the ControlNet.

Style Guidance. We employ a style adapter consisting of a text encoder and an image encoder [42], integrated into the cross-attention layers for style guidance extraction and injection. The text prompt is fixed as "a photo of a portrait," while image features is extracted from the penultimate layer of CLIP. The extracted latent representation is then passed through a projection layer to obtain the style guidance $c^{sty} = [c^t, c^i]$, where c^t represents the text feature and c^i represents the image feature. These features are then fused with the main Denoising U-Net. Specifically, we perform decoupled cross-attention operations on both text and image features, injecting them into the corresponding layers of the Denoising U-Net. The decoupled cross-attention mechanism can be illustrated as:

$$Z^{new} = \text{softmax}\left(\frac{QK^t}{\sqrt{d}}\right)V^t + \lambda \cdot \text{softmax}\left(\frac{QK^i}{\sqrt{d}}\right)V^i, \qquad (9)$$

where λ is a weight that balances the contributions of text and image prompts. Q, K^t, V^t represent the query, key, and value matrices for text cross-attention, while K^i and V^i correspond to image cross-attention. Given the query features Z from the main denoising UNet, we define $Q =$

ZW_q, $K^i = c^iW_k^i$, $V^i = c^iW_v^i$, $K^t = c^tW_k^t$ and $V^t = c^tW_v^i$, where only W_k^i and W_v^i are trainable weights.

4.3. Initial Latent AdaIN-Wavelet Transform

In portrait style transfer, color tone plays a crucial role in conveying style information. If sampling begins from the latent z_T^c of the content image, the generated image will largely retain its original color tone. To achieve an initial latent representation that preserves the structural details of the input image while enhancing local color style transfer, we propose a novel AdaIN-Wavelet transform for latent initialization. First, we compute the latent representation of the warped reference image $z_0^{s\text{-}w}$ using DDIM inversion, denoted as $z_T^{s\text{-}w}$. Since initiating the denoising process from $z_T^{s\text{-}w}$ strongly enhances color transfer, directly using it may result in a blurred output. To address this, we propose blending the low-frequency components of $z_T^{s\text{-}w}$ with the high-frequency components of the content latent, striking a balance between effective stylization and content detail preservation. We first apply AdaIN between $z_T^{s\text{-}w}$ and z_T^c, formulated as:

$$z_T^{cs'} = \sigma(z_T^{s\text{-}w})\left(\frac{z_T^c - \mu(z_T^c)}{\sigma(z_T^c)}\right) + \mu(z_T^{s\text{-}w}), \quad (10)$$

where $\sigma()$ and $\mu()$ denote channel-wise mean and standard deviation, respectively. We then apply the wavelet transform to integrate the high-frequency information from $z_T^{cs'}$ and the low-frequency informatoin from $z_T^{s\text{-}w}$, which is obtained as:

$$z_T^{cs} = \text{IDWT}(z_{T,LL}^{s\text{-}w}, z_{T,LH}^{cs'}, z_{T,HL}^{cs'}, z_{T,HH}^{cs'}), \quad (11)$$

where IDWT denotes the inverse wavelet transform.

To control the strength of stylization, we introduce a parameter γ. Since stylization is influenced by both the latent initialization and the style adapter, γ is applied in two key areas. First, we interpolate between the input latent z_T^c and stylized latent z_T^{cs}:

$$z_T^{cs} = \gamma \times z_T^{cs} + (1 - \gamma) \times z_T^c. \quad (12)$$

Similarly, γ is used in style adapter to interpolate between the feature embeddings of the warped reference image and the input image. In this paper, we set $\gamma = 1.0$ by default for best stylization effects.

4.4. Training and Implementation Details

We implement our method using Stable Diffusion 1.5 and follow a two-stage training strategy. In the first stage, we train the semantic adapter while freezing the other parameters of the pre-trained diffusion model. To ensure the features encode meaningful information, we incorporate a noise prediction loss during this stage:

$$\mathcal{L}_{sem} = \mathbb{E}_{z_t,t,I^x,\epsilon \sim \mathcal{N}(0,I)}\|\epsilon - \epsilon_\theta(z_t, t, I^x)\|_2^2, \quad (13)$$

Method	Gram loss↓	LPIPS↓	ID↓
Shih et al. [28]	0.376	0.187	0.093
Chen et al. [1]	0.287	0.156	0.344
Zhou et al. [48]	0.688	0.393	0.657
Wang et al. [38]	**0.208**	0.181	0.106
IP-A[42] + C.N.[44]	2.835	0.245	0.774
Deng et al. [5]	1.745	0.162	0.393
StyleID[3]	0.505	0.198	0.222
InstantStyle + [36]	0.557	0.294	0.272
Ours	0.274	**0.116**	**0.057**
Ours (w/o ControlNet)	0.236	0.333	0.450
Ours (w/o style adapter)	0.548	0.145	0.086
Ours (w/ Init AdaIN)	1.196	0.151	0.062

Table 1. **Quantitative comparison on CelebAMask-HQ.**

Method	Gram loss↓	LPIPS↓	ID↓
Shih et al. [28]	0.802	0.156	0.105
Chen et al. [1]	0.926	0.169	0.316
Zhou et al. [48]	0.827	0.423	0.669
Wang et al. [38]	1.488	0.119	0.096
IP-A[42] + C.N.[44]	4.211	0.375	0.763
Deng et al. [5]	1.493	0.178	0.331
StyleID[3]	1.343	0.149	0.165
InstantStyle+ [36]	0.723	0.192	0.203
Ours	**0.657**	**0.083**	**0.087**

Table 2. **Quantitative comparison on a mixed dataset with portraits from multiple domains.**

where I^x represent an arbitrary input image. The total loss at this stage is:

$$\mathcal{L} = \mathcal{L}_{sem} + \lambda_c\mathcal{L}_{cyc} + \lambda_m\mathcal{L}_{mask}, \quad (14)$$

where λ_c and λ_m are set to 1 and 10, respectively.

In the second training stage, we train the ControlNet and style adapter. Due to the absence of paired training data for our tasks, we train the model to reconstruct images conditioned on both the style and content inputs. The loss function at this stage can be formulated as follows:

$$\mathcal{L}_{rec} = \mathbb{E}_{z_t,t,c^{cnt},c^{sty},\epsilon \sim \mathcal{N}(0,I)}\|\epsilon - \epsilon_\theta(z_t, t, c^{cnt}, c^{sty})\|_2^2, \quad (15)$$

where c^{cnt} and c^{sty} are both extracted from a single input image.

Our method is implemented in PyTorch [23]. We utilize 2 Nvidia RTX 4090 GPUs with a batch size of 1 per GPU for training. The AdamW [20] optimizer is employed, with a learning rate of 1e-5. The first stage of training runs for 500K iterations, while the second stage is trained for 300K iterations. The training process requires approximately 24GB of memory, and testing necessitates 16GB of
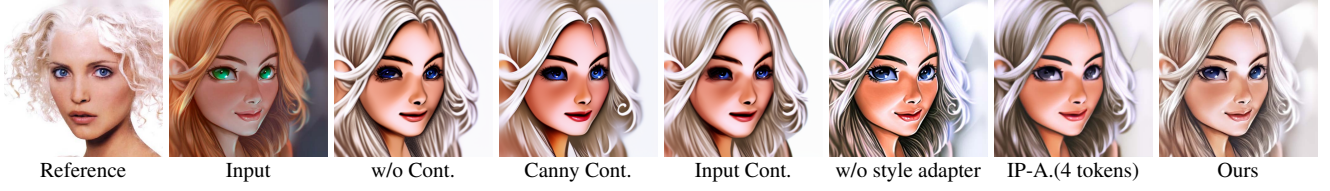
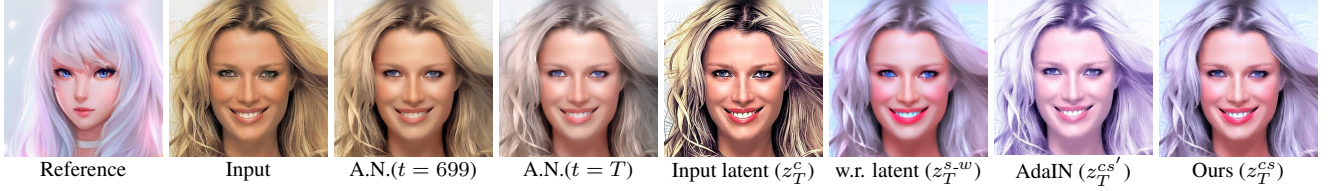Figure 5. **Effect of the ControlNet and style adapter.**

Figure 6. **Effect of the initialization of latent noise.** "A.N." means initializing latent by adding noise of level t to input. "w.r." stands for warped reference.
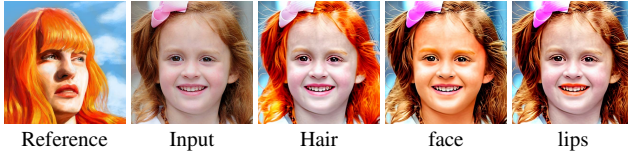

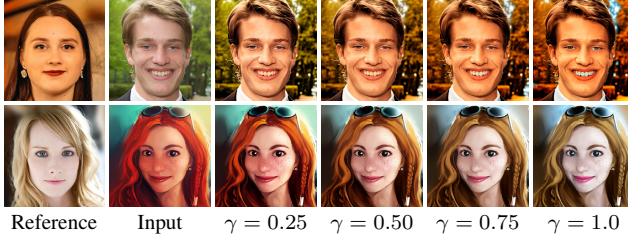Figure 7. **Controllable region-specific style transfer.**

Figure 8. **Style interpolation.**

memory for images with a resolution of 512×512. For testing, we use the DDIM sampler [18] with 10 timesteps for inversion and 30 timesteps for sampling.

5. Experiments

5.1. Datasets and Evaluation Metrics

Datasets. We evaluate our method on CelebAMask-HQ [13], FFHQ [11], and AAHQ [19]. CelebAMask-HQ contains 30K high-resolution portrait images, each paired with a facial segmentation mask covering 19 facial semantic categories. FFHQ consists of 70K high-quality images depicting a wide range of human faces. AAHQ includes approximately 25K high-quality artistic images with varied painting styles, color tones, and face attributes. To demonstrate the generalization ability of our method, we train it solely on the 28K training images from the CelebAMask-HQ dataset [38].

Metrics. To evaluate the stylization effects for semantically

corresponding regions, we follow Wang et al. [38] to compute the Gram loss between the input image and the warped reference. Additionally, we use LPIPS [45] to measure content preservation performance and adopt the Identity Distance (ID) [4], as used in [14, 22], to quantitatively assess identity preservation.

5.2. Comparison with State-of-the-art Methods

Baselines. We compare our method with two types of state-of-the-art style transfer methods: (i) conventional portrait style transfer methods, including Shih et al.[28], Chen et al. [1], Zhou et al. [48], and Wang et al. [38], and (ii) diffusion-based style transfer methods, including IP-Adapter (IP-A) [42] + ControlNet (C.N.) [44], Deng et al. [5], StyleID [3], and InstantStyle-Plus (InstantStyle+) [36]. For learning-based methods, we retrain them on CelebAMask-HQ using the publicly available training code provided by the authors, with the recommended parameter settings. For methods with pre-trained models only, we fine-tune their models on CelebAMask-HQ.

Quantitative Evaluation. Tables 1 and 2 present the quantitative comparison results on the CelebAMask-HQ dataset and a mixed dataset from multiple domains, including portraits from CelebAMask-HQ, FFHQ, and AAHQ for a more comprehensive evaluation. As shown, on CelebAMask-HQ, the Gram loss of our method is comparable to the state-of-the-art method [38], while our LPIPS and ID scores are significantly better, demonstrating superior content and identity preservation. On the mixed dataset, our method achieves the lowest Gram loss. Furthermore, our method outperforms other style transfer techniques in both LPIPS and ID, highlighting its strong domain generalizability.

Qualitative Evaluation. Figure 3 compares our results with those of previous style transfer methods [3, 5, 38, 42, 44]. Our method achieves more faithful stylization
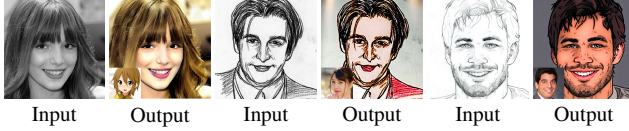
Figure 9. **Results of adding color to grayscale and sketch portraits.**
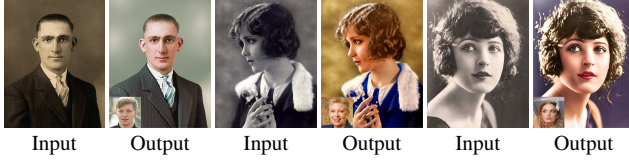

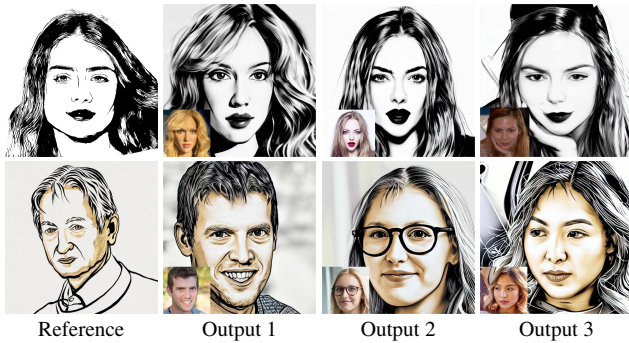Figure 10. **Results of modernizing old photographs.**

Figure 11. **Results of transferring real portraits to sketch style.**

effects by effectively transferring styles across semantically aligned regions, such as skin, lips, hair, and background. Additionally, our method demonstrates superior performance in content preservation, unlike artistic methods [5, 42, 44], which often alter facial identity. Please see the supplementary material for more visual results.

5.3. More Analysis

Ablation Studies. Figure 5 investigates the effects of ControlNet and the style adapter. Our high-frequency ControlNet preserves content details more effectively than using a canny edge detection map or the input image. Moreover, with the style adapter, the accuracy of stylization is further enhanced, surpassing the performance of the common IP-Adapter with 4 tokens. Figure 6 compares results using different initializations of latent noise. Initializing latent by adding noise of different levels to input, or using the inversion input latent as the initial tend to preserve the original color style. Performing AdaIN between the input and reference latents [3] improves overall stylization effects but doesn't ensure local style similarity. Using the latent of the warped reference yields strong stylization effects, but the output becomes blurry. In comparison, our AdaIN-Wavelet initialization strikes a better balance between stylization and content detail preservation. Quantitative results in Table 1

Metric	Deng et al.	StyleID	InstantStyle+	Ours
Time ↓ (sec)	24.18	9.12	67.40	**6.97**

Table 3. **Comparison on inference time.**

further support this conclusion. More results can be found in the supplementary material.

Controllable Region-specific Style Transfer. As shown in Figure 7, our method generates natural results by transferring the style of specific semantic regions, such as hair, face, or lips. Given the mask of a specific region, we achieve this by simply replacing the corresponding regions in the warped reference with those from the input portrait.

Style interpolation. Figure 8 demonstrates that our method supports continuous style interpolation, allowing for control over the stylization strength by adjusting the parameter γ as described in Section 4.3.

Portrait Style Transfer across Different Domains. Figure 9 demonstrates that our method can colorize grayscale and sketch portraits using reference portraits from different domains. Figure 10 showcases the ability of our method to restore the color of old photos. Finally, Figure 11 illustrates that our method can replicate a sketch style. Note, by adjusting the ControlNet conditioning scale, our method allows for controlling the level of detail preservation for better stylization effects. Please see the supplementary material for more visual results.

Inference time. Our model processes images at a resolution of 512×512 in approximately 6.97 seconds on an NVIDIA RTX 4090 GPU, making it more efficient than other diffusion-based methods, as shown in Table 3. The efficiency comes from two factors: (i) our AdaIN-Wavelet module is robust to the number of DDIM inversion steps—only 10 steps are needed to achieve high-quality results; (2) our content guidance is delivered via ControlNet, which helps remove the necessity for intermediate feature extraction [3, 5]. These advantages collectively contribute to the accelerated inference.

6. Conclusion

We have presented a portrait style transfer framework that excels in domain generalization and produces high-quality results. By leveraging pre-trained diffusion models and a semantic adapter, we get reliable dense correspondence between portraits, enabling semantically aligned stylization. Besides, we develop an AdaIN-Wavelet transform to balance stylization and content preservation, and also introduce a style adapter as well as a ControlNet to offer style and content guidance. Extensive experiments demonstrate that our method outperforms existing approaches.

References

[1] Shu-Yu Chen, Feng-Lin Liu, Yu-Kun Lai, Paul L Rosin, Chunpeng Li, Hongbo Fu, and Lin Gao. Deepfaceediting: deep face generation and editing with disentangled geometry and appearance control. *ACM Transactions on Graphics (TOG)*, 40(4):1–15, 2021. 2, 3, 6, 7

[2] Hansam Cho, Jonghyun Lee, Seunggyu Chang, and Yonghyun Jeong. One-shot structure-aware stylized image synthesis. In *CVPR*, pages 8302–8311, 2024. 2

[3] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *CVPR*, 2024. 2, 6, 7, 8, 1

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 7

[5] Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. z^*: Zero-shot style transfer via attention rearrangement. In *CVPR*, 2024. 2, 6, 7, 8, 1

[6] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 2

[7] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *CVPR*, 2019. 2, 3

[8] Meiqin Guo and Jianmin Jiang. A robust deep style transfer for headshot portraits. *Neurocomputing*, 361:164–172, 2019. 3

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeuIPS*, 33:6840–6851, 2020. 3

[10] Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. Training-free content injection using h-space in diffusion models. In *WACV*, pages 5151–5161, 2024. 2

[11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3, 7

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

[13] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 7

[14] Xiaoming Li, Xinyu Hou, and Chen Change Loy. When stylegan meets stable diffusion: a $\mathcal{W}_+$ adapter for personalized image generation. In *CVPR*, 2024. 7

[15] Xinghui Li, Jingyi Lu, Kai Han, and Victor Adrian Prisacariu. Sd4match: Learning to prompt stable diffusion model for semantic matching. In *CVPR*, pages 27558–27568, 2024. 2

[16] Zonglin Li, Xiaoqian Lv, Wei Yu, Qinglin Liu, Jingbo Lin, and Shengping Zhang. Face shape transfer via semantic warping. *Visual Intelligence*, 2, 2024. 2

[17] Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. Ctrl-x: Controlling structure and appearance for text-to-image generation without guidance. *arXiv preprint arXiv:2406.07540*, 2024. 2

[18] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 7

[19] Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. Blendgan: Implicitly gan blending for arbitrary stylized face generation. In *NeurIPS*, 2021. 7

[20] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[21] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *CVPR*, 2017. 2

[22] Jiteng Mu, Shalini De Mello, Zhiding Yu, Nuno Vasconcelos, Xiaolong Wang, Jan Kautz, and Sifei Liu. Coordgan: Self-supervised dense correspondences emerge from gans. In *CVPR*, 2022. 7

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 6

[24] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *CVPR*, pages 8693–8702, 2024. 2

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICLR*, pages 8748–8763, 2021. 3

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3

[27] Ahmed Selim, Mohamed Elgharib, and Linda Doyle. Painting style transfer for head portraits using convolutional neural networks. *ACM Transactions on Graphics (ToG)*, 35(4):1–18, 2016. 3

[28] Yichang Shih, Sylvain Paris, Connelly Barnes, William T. Freeman, and Frédo Durand. Style transfer for headshot portraits. *ACM Transactions on Graphics (TOG)*, 33:1 – 14, 2014. 2, 6, 7

[29] Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. Portrait lighting transfer using a mass transport approach. *ACM Trans. Graph.*, 36 (4), 2017. 2

[30] Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, and Tat-Jen Cham. Agilegan: Stylizing portraits by inversion-consistent transfer learning. *ACM Trans. Graph.*, 40(4), 2021. 3

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[32] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022. 2

[33] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *NeurIPS*, 36:1363–1389, 2023. 2, 4, 5

[34] Cong Wang, Fan Tang, Yong Zhang, Tieru Wu, and Weiming Dong. Towards harmonized regional style transfer and manipulation for facial images. *Computational Visual Media*, 9 (2):351–366, 2023. 2, 3

[35] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 2

[36] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024. 2, 5, 6, 7, 1

[37] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2

[38] Xinbo Wang, Qing Zhang, Yongwei Nie, and Wei-Shi Zheng. Towards photorealistic portrait style transfer in unconstrained conditions. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–14, 2025. 2, 6, 7

[39] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *ICCV*, pages 7677–7689, 2023. 2

[40] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Vtoonify: Controllable high-resolution portrait video style transfer. *ACM Trans. Graph.*, 41(6), 2022. 3

[41] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *CVPR*, 2022. 2, 3

[42] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 4, 5, 6, 7, 8

[43] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *NeurIPS*, 36, 2024. 2, 4

[44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 6, 7, 8

[45] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[46] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023. 2

[47] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *CVPR*, pages 10146–10156, 2023. 2

[48] Yang Zhou, Zichong Chen, and Hui Huang. Deformable one-shot face stylization via dino semantic guidance. In *CVPR*, pages 7787–7796, 2024. 2, 3, 6, 7