One-shot Portrait Stylization via Geometric Alignment

Xinrui Wang[1] Zilin Guo[2] Zhuoru Li[3] Jinze Yu[4]
Heng Zhang[5] Yusuke Iwasawa[1] Yutaka Matsuo[1] Jiaxian Guo[1]
[1]The University of Tokyo [2]Xiaomi Corporation [3]Project HAT
[4]Waseda University [5]Adaspace Technology Co.Ltd

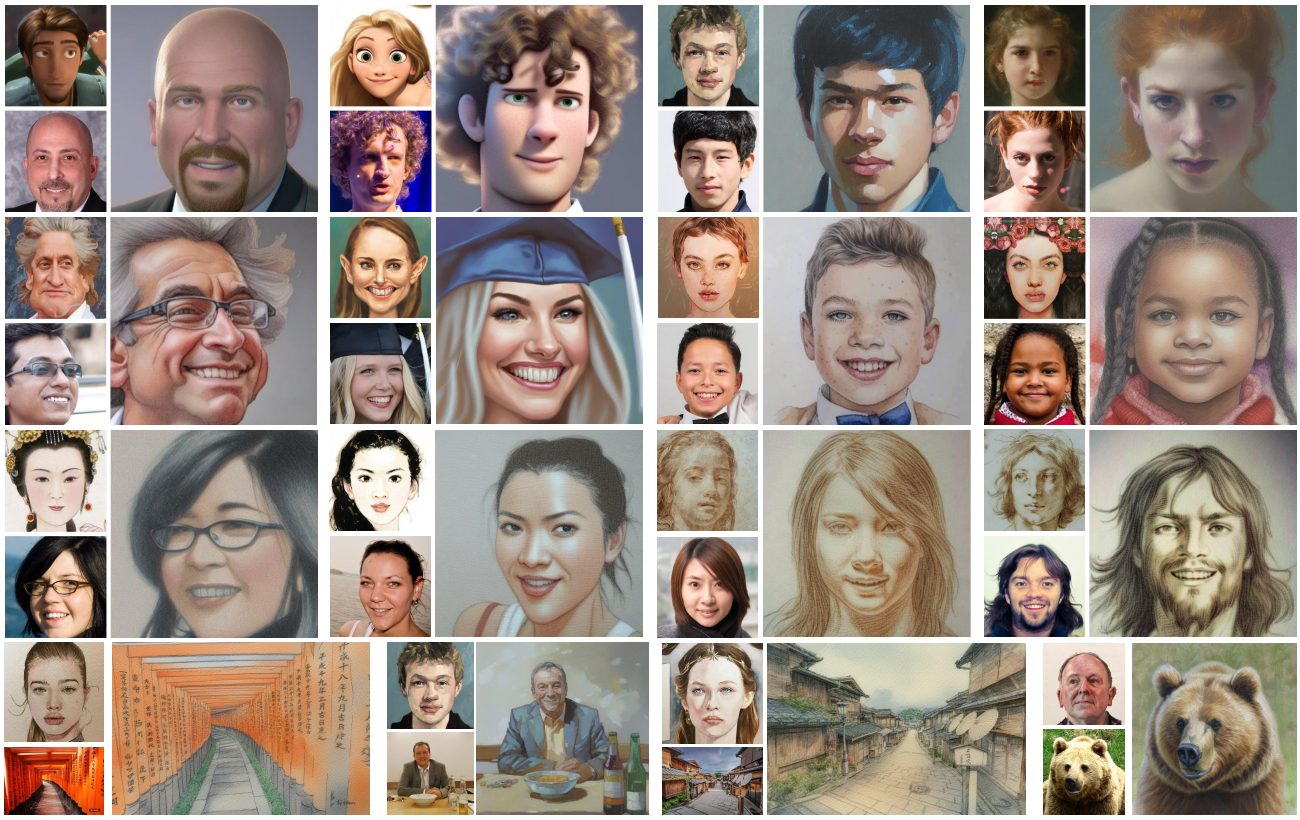

Figure 1. Given a single style reference, our method can be effectively trained with unpaired portrait images to learn style information to synthesize high quality stylized portrait images, and even images not in training data, such as full body photos, animals and scenery.

Abstract

Portrait stylization casts vivid artistic style drawn from style examples to portrait photos. Although recently extensively studied with machine learning algorithms, existing methods still face challenges in stylizing portraits from a single style reference, severely limiting their potential for real-world applications. In this paper, we propose a portrait stylization method that learns style reference from a single artistic portrait image. Unlike previous StyleGAN based methods that heavily rely on the quality of GAN inversion or diffusion based methods that introduce computational expensive operations and fall short of precise control, our method achieves high-quality stylization with small computation and parameter budget. Specifically, we employ geometric alignment to build spatial correlation between content images and style reference. A geometry LoRA and a style LoRA are then jointly optimized based on a pre-trained diffusion backbone respectively, with orthogonal adaptation used to disentangle the geometry and style information. During inference, the style LoRA is integrated into the diffusion backbone and ControlNet is further combined to facilitate better spatial and identity control. We illustrate abundant stylized portraits with multiple styles. Qualitative comparison, quantitative validation and user study prove that our method outperforms existing methods, and ablation study demonstrates the effectiveness of each components.

1. Introduction

Centuries ago, artists created immortal artworks such as *Mona lisa* and *Girl with a Pearl Earring* with pigments and canvas. Today, everyone can apply artistic styles to their selfies with camera filters and post them on social media. The creation of artwork is often inspired by an existing piece of art. For example, *Vincent Van Gogh* created a series of works based on Japanese Ukiyoe style. Can we also create artistic portraits with a given style reference? This long-standing question makes one-shot portrait stylization a highly demanding task.

Recently, machine learning methods have been widely explored for one-shot portrait stylization. StyelGAN and its derivatives [3, 15, 16] achieved good performance by employing StyleGAN inversion and style mixing to generate paired dataset from a single style reference and fine-tune a generator network for one-shot stylization. However, synthesizing and manual cleaning paired datasets is time consuming and labor intensive, and data quality is susceptible to the inversion algorithm. Diffusion models [9, 24] have since emerged as powerful tools for image synthesis and also widely adopted for stylization. InstantStyle [30] relies on empirical attention adjustments, inherently limiting precise control due to its adapter-based design. OSASIS [2] incurs a significant training burden by fine-tuning the entire model with auxiliary modules. Furthermore, Pair-customization [14]'s reliance on paired images compromises its practical applicability.

To address these issues, we start from our observation that the persistent challenge of portrait stylization lies in the vast diversity in the geometry of human faces. Additionally, style references often feature exaggerated artistic effects, leading to significant discrepancies in the position and shape of facial characteristics compared to actual human portraits. These geometric gaps make it difficult for models to disentangle the identity information and the style information of the style reference, deteriorating the stylization process. Specifically, we propose to combine geometric alignment with the powerful Stable Diffusion backbone [24] and LoRA (Low Rank Adaption) [11] optimization to facilitate stylization. Given a single artistic portrait image as the style reference, we employ Thin Plate Spline (TPS) to warp each randomly sampled portrait image for training and the style reference to the average facial landmarks to eliminate the geometry gap and simplify the following stylization process. Orthogonal adaptation [21] is then employed to disentangle the style information and the geometry information, enabling corresponding LoRA weights to be jointly optimized. During inference, the trained style LoRA is merged into the diffusion backbone to synthesize stylized images, and ControlNet [39] is integrated into the pipeline to provide spatial guidance and preserve spatial and identity information for portrait stylization.

Experiments show that our method can be trained for only a few hundred steps with a single style reference, and synthesize high-quality portrait images loyally representing the referenced artistic styles. Our results show superior synthesis quality and fidelity, as well as better identity preservation over existing methods in the qualitative comparison, and also achieve better performance in quantitative evaluation. In user study, our method is also subjectively preferred by most candidates. The ablation study demonstrates the effectiveness of each component in our framework.

In summary, our contributions are as follows: 1.We introduce a geometric alignment mechanism to bridge the domain gap between style examples and portrait images, facilitating robust geometry-style disentanglement. 2. We jointly optimize geometry LoRA and Style LoRA modules built upon a diffusion backbone to extract respective information. This forms a novel and effective portrait stylization framework. 3. Through qualitative evaluation, quantitative comparison, and user study, we demonstrate that our proposed method generates high-quality results and outperforms existing works. Code, dataset, and trained weights will be publicly available upon acceptance of the paper.

2. Related Works

2.1. Non-photorealistic Rendering

Non-photorealistic rendering (NPR) transforms photographic images into artistic styles. Early approaches relied on iterative optimization to simulate artistic textures such as pencil drawings [19, 36], oil paintings [8], and example-based 2D stylizations [29]. The advent of deep learning, particularly Convolutional Neural Networks (CNNs) [5, 13], significantly advanced stylization quality and efficiency. Image-to-image (I2I) translation frameworks [12, 13, 32, 42] became prominent in NPR, employing trained transformation networks for flexible inter-domain stylization, and had been applied on various styles [33, 37, 38, 40].

While stylization can introduce distortions, incorporating spatial alignment has been explored to improve output quality. For example, [1] integrated spatial relation-augmented modules into VGG-based architectures for structural coherence , [6] performed rigid alignment in feature space by treating feature maps as point clouds. Nevertheless, these techniques primarily exploit spatial information implicitly via unsupervised attention mechanisms. This lack of explicit semantic supervision limits their effectiveness, thereby constraining the fidelity and controllability of stylization outcomes.

In this paper, we employ the semantic information inherent in highly structured human faces to enhance the stylization process. Geometric alignment is explicitly applied to facial images and the style example through the use of differentiable TPS transformations to align them on the same

geometric position. This enables the geometry information and style information in the style example to be disentangled and learned by the joint optimization of two LoRA weights, realizing single shot portrait stylization with low computational budget.

2.2. Diffusion Models

Diffusion Probabilistic Models [9] are a class of generative models inspired by nonequilibrium thermodynamics [27] and have achieved great success in image synthesis and editing. The denoising process is learned by reversing the forward diffusion process $_0, \ldots, _T$, where image $_0$ is progressively diffused to random noise $_T$ over T timesteps following $_t = \sqrt{\bar{\alpha}_t}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ for timestep $t \in [0, T]$. Noise $\epsilon \sim \mathcal{N}(0, I)$ is randomly sampled, and $\bar{\alpha}_t$ controls the noise strength. The training objective is to denoise intermediate noisy image $_t$ via noise prediction:

$$\mathbb{E}_{\epsilon,x,c,t}\left[w_t \|\epsilon - \epsilon_\theta(x_t, c, t)\|^2\right], \tag{1}$$

where w_t represents a time-dependent weight, $\epsilon_\theta(\cdot)$ represents the denoiser that learns to predict noise, and $\mathbf{c}$ denotes the conditioning inputs, such as text prompts. In inference, the denoiser ϵ_θ gradually denoises random Gaussian noise into images. The model is trained to generate images that approximate the distribution of training data [9]. In this work, Stable Diffusion XL [22], a large-scale text-to-image diffusion model built on Latent Diffusion Models [24], is used as the diffusion backbone.

2.3. Portrait Stylization

Portrait stylization is a persistent problem in computer vision, driven by its artistic and practical value. Traditional exemplar-based methods [7, 26, 28] typically follow an align-and-transfer pipeline, where local statistics from a style exemplar are mapped to a target portrait. However, this approach lacks generalizability across diverse styles, as it does not learn from a dataset. Recent StyleGAN based approach [3] exploits its powerful generative capabilities, yet the results are highly dependent on precise inversion and often suffer from artifacts. Diffusion-based methods [2, 4, 30, 31] excel at high-quality synthesis, but they face challenges in maintaining consistency from a single style reference. Yet a critical limitation of these approaches is their frequent failure to preserve spatial structure and identity, which are paramount in portrait stylization.

In this work, we propose a novel approach that leverages a powerful pre-trained diffusion backbone and integrates geometric alignment with disentangled representations to optimize Low-Rank Adaptation (LoRA) weights for stylization. Our method enables efficient one-shot stylization, requiring only a single style reference and minimal computational resources. Specifically, stylization can be achieved with low computational budget involving just a few hundred steps and a small set of trainable LoRA parameters.

3. Method

Given a random artistic portrait as a style reference, our goal is to learn a model that captures its style and transfers them to arbitrary portrait images. To achieve this, we first leverage readily accessible facial landmarks from both the style reference and randomly sampled unpaired training portrait images to establish geometric alignment between them. Subsequently, we apply orthogonal adaptation to a pretrained Stable Diffusion XL (SDXL) model [22] to disentangle geometry and style representations. Specifically, we jointly optimize a geometry-specific LoRA and a style-specific LoRA with the geometrically aligned data.

In the inference phrase, the optimized style LoRA is integrated into the diffusion backbone and employed as a replacement for the standard classifier-free guidance to steer the generation toward the desired style. Using the same random seed, we can synthesize paired outputs: a photo-realistic portrait generated by the base SDXL model and its stylized counterpart produced by the SDXL model with the style LoRA integrated, as illustrated in Figure 4.

To stylize existing portrait images, we incorporate the style LoRA with SDXL backbone and pair it with ControlNet for image-to-image translation. We provide an overview of the training pipeline in Figure 2 and the inference process in Figure 3. Detailed descriptions of each component are presented in the subsequent sections.

3.1. Geometric Alignment

Learning style information from a single reference image is challenging and easily leads to overfitting [4,25]. Particularly, training a stylization LoRA and employ it on the same domain can easily result in copying content [14]. In the proposed method, a random artistic portrait image is given as the style reference, which poses additional challenges to the model to distinguish style information and geometry information from non-relevant content-style pairs. To address this issue, we propose to geometrically align content images and the style example to simplify the stylization process.

In the training phrase, the style reference is paired with different content images randomly sampled for each training step. For each content-style image pair, we extract the facial landmarks of both images. Affine transformation is at first used to align them, then Thin Plate Spline (TPS) transformation is employed to warp both images to the mean landmarks.We show in the Experiments section that, instead of warping one image to the landmarks of another, warping both images to the mean landmark positions guarantees high-quality warping and thus enhances stylization quality.

Equation 2 shows the TPS based geometric alignment, where I^c and I^s represent content image and style image,

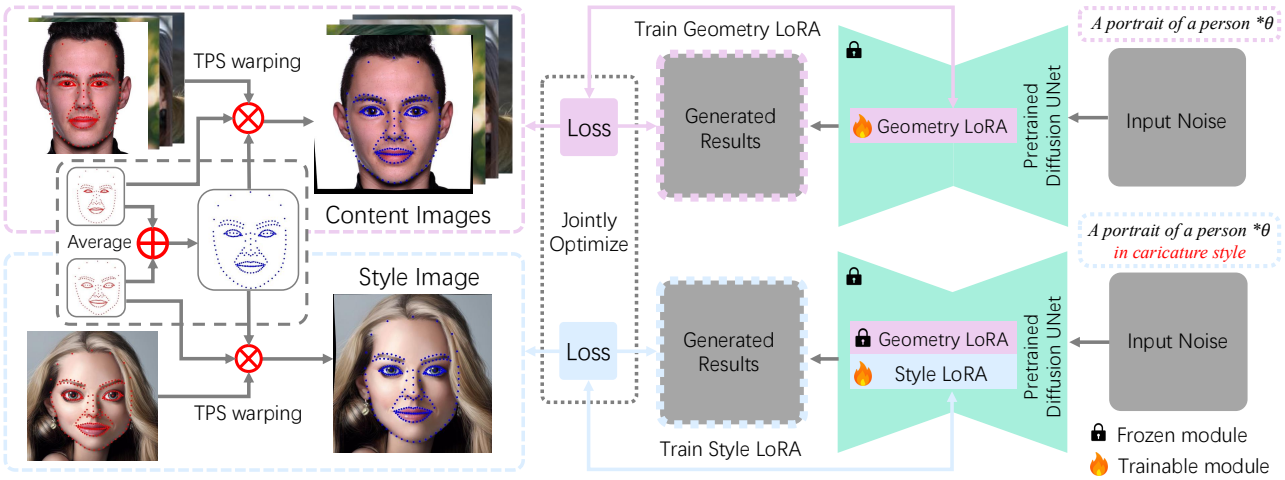

Figure 2. Training process of the proposed framework. We extract the facial landmarks of content image and style example and warp them to the mean geometric shape, then jointly optimized a geometry LoRA and a style LoRA with orthogonal adaptation for disentanglement.

and P^c and P^s represent the landmarks of content image and style image respectively.

$$\begin{aligned} I^c_{align} &= \mathcal{TPS}(I^c, P^c, P^{mean}) \\ I^s_{align} &= \mathcal{TPS}(I^s, P^s, P^{mean}) \\ \text{s.t.} P^{mean}_i &= 0.5 * P^c_i + 0.5 * P^s_i, \quad i = 1, 2, \ldots, N \end{aligned} \tag{2}$$

3.2. Stylization with Orthogonality Adaption

Geometric alignment helps establish spatial correspondence between content-style image pairs, facilitating the model's learning of stylistic attributes from image pairs. To better extract style information from the reference, we propose to disentangle the geometry information and style information from the style example and jointly train a geometry LoRA and a style LoRA. The geometry LoRA is tasked with reconstructing the portrait geometry, as the content images and the style reference are warped to identical geometry represented by landmarks. The style LoRA is combined with geometry LoRA and utilized for synthesizing the stylized image.

During the training phase, the geometry LoRA is conditioned on a geometry-specific text prompt "T_{geometry}" augmented by a random rare token $*\theta$. The combined LoRAs are guided by a style text prompt, formed by appending a style suffix, such as "in disc style" (e.g., "in oil paint style" or "in sketch style"), to the geometry-specific text.

Following [21], we further exert orthogonality adaption to facilitate the disentanglement of geometry and style information for better stylization. For each layer in the original diffusion model with an initial weight $W_0 \in \mathbb{R}^{m\times n}$, we use $W_{geometry}$ and W_{style} to denote the modified weight of geometry LoRA and style LoRA respective, and both LoRAs are used to reconstruct the image. The LoRA learning process is formulated as:

$$\begin{aligned} W_{\text{geometry}} &= W_0 + B_{\text{geometry}} A_{\text{geometry}} \\ W_{\text{style}} &= W_0 + B_{\text{style}} A_{\text{style}} \\ W_{\text{combined}} &= W_0 + B_{\text{geometry}} A_{\text{geometry}} + B_{\text{style}} A_{\text{style}} \end{aligned} \tag{3}$$

where $B \in \mathbb{R}^{m\times r}$, $A \in \mathbb{R}^{r\times n}$, and $r \ll min(m, n)$. Especially, B_{geometry} and B_{style} are initialized from zero matrix and A_{geometry} and A_{style} are from an orthonormal basis. During training, A_{geometry} and A_{style} are fixed and only B_{geometry} and B_{style} are updated. This enforces the geometry LoRA and style LoRA to be optimized responding to orthogonal inputs, facilitate the disentanglement of geometry and style information to improve the quality of stylization.

To optimize two different LoRA weights, we also propose two loss fuctions as learning objectives. When reconstructing the content image, we adopt the standard training objective for diffusion models to optimize the geometry LoRA:

$$\mathcal{L}_{\text{geometry}} = \mathbb{E}_{\epsilon, \boldsymbol{x}_{\text{geometry}}, t}\left[w_t \|\epsilon - \epsilon_{\theta_{\text{geometry}}}(\boldsymbol{x}_{t,\text{geometry}}, \boldsymbol{p}_{\text{geometry}}, t)\|^2\right] \tag{4}$$

where $\epsilon_{\theta_{\text{geometry}}}$ represents the denoising model with geometry LoRA integrated, $\boldsymbol{x}_{t,\text{geometry}}$ represents a noisy content image at timestep t, and $\boldsymbol{p}_{\text{geometry}}$ represents the geometry-specific text prompt with the rare token $*\theta$.

When synthesizing the style image, we use the combined weight with both geometry LoRA and style LoRA integrated into the model. During training, the weights of geometry LoRA are fixed and we only update the weights of style LoRA with the following formula:

$$\mathcal{L}_{\text{combined}} = \mathbb{E}_{\epsilon, \boldsymbol{x}_{\text{style}}, t}\left[w_t \|\epsilon - \epsilon_{\theta_{\text{combined}}}(\boldsymbol{x}_{t,\text{style}}, \boldsymbol{p}_{\text{style}}, t)\|^2\right] \tag{5}$$

where $\epsilon_{\theta_{\text{combined}}}$ represents the denoising model with both geometry LoRA and style LoRA integrated, and $\boldsymbol{p}_{\text{style}}$ represents the text prompt "$\{\boldsymbol{p}_{\text{geometry}}\}$ in $\{$disc$\}$ style", with

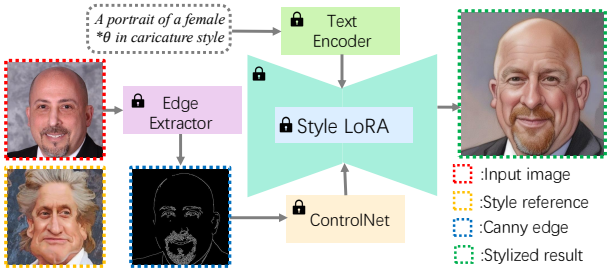

Figure 3. Image-to-image translation inference for facial stylization. ControlNet with canny edge map is utilized as control condition to provide ID information.

{disc} describing the specific style (e.g., "oil paint"). The LoRA weights are jointly optimized by the combination of the two losses:

$$\min_{\Delta\theta_{\text{geometry}},\Delta\theta_{\text{style}}} \mathcal{L}_{\text{geometry}} + \mathcal{L}_{\text{combined}} \tag{6}$$

3.3. Style Guided Image Stylization

Classifier-free guidance (CFG) [10] is a common technique to improve the synthesis quality of text-to-image model. We use $\hat{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{p}, t)$ to represent the new noise prediction, $\varnothing$ denotes no conditioning, and λ_{cfg} controls the amplification of text guidance, and omitting the timestep t for notation simplicity, then the CFG is formulated as:

$$\hat{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{p}) = \epsilon_\theta(\boldsymbol{x}_t, \varnothing) + \lambda_{\text{cfg}}(\epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{p}) - \epsilon_\theta(\boldsymbol{x}_t, \varnothing)) \tag{7}$$

To preserve the content information and improve the quality of stylized results, we propose to add weighted style guidance while preserving the original denoising path during inference, where style guidance is referred as the difference in noise prediction between style LoRA and the pre-trained model:

$$\begin{aligned}\hat{\epsilon}_{\theta_0,\theta_{\text{style}}}(\boldsymbol{x}_t, \boldsymbol{p}, \boldsymbol{p}_{\text{style}}) = \epsilon_{\theta_0}(\boldsymbol{x}_t, \varnothing) \\ + \lambda_{\text{cfg}}(\epsilon_{\theta_0}(\boldsymbol{x}_t, \boldsymbol{p}) - \epsilon_{\theta_0}(\boldsymbol{x}_t, \varnothing)) \\ + \lambda_{\text{style}}(\epsilon_{\theta_{\text{style}}}(\boldsymbol{x}_t, \boldsymbol{p}_{\text{style}}) - \epsilon_{\theta_0}(\boldsymbol{x}_t, \boldsymbol{p})),\end{aligned} \tag{8}$$

λ_{style} is used to control the style guidance strength. When $\lambda_{\text{style}} = 0$, it is equivalent to generating original content.

During inference, we can generate paired photo-realistic and stylized portraits by using a fixed random seed to ensure correspondence. The photo-realistic portrait is synthesized by the SDXL backbone conditioned on a baseline text prompt that encodes the content.To obtain the stylized counterpart, we integrate the trained style LoRA into the SDXL backbone and use the same baseline prompt augmented with a style description, thereby guiding the generation toward the desired artistic appearance. We show the results of synthesized image pairs in Figure 4

To enable stylization of existing portrait images, we further incorporate ControlNet [39] guided by Canny edge maps into our pipeline to perform image-to-image translation. As illustrated in Figure 3, this design effectively produces high-quality stylization results while faithfully preserving the subject's identity.

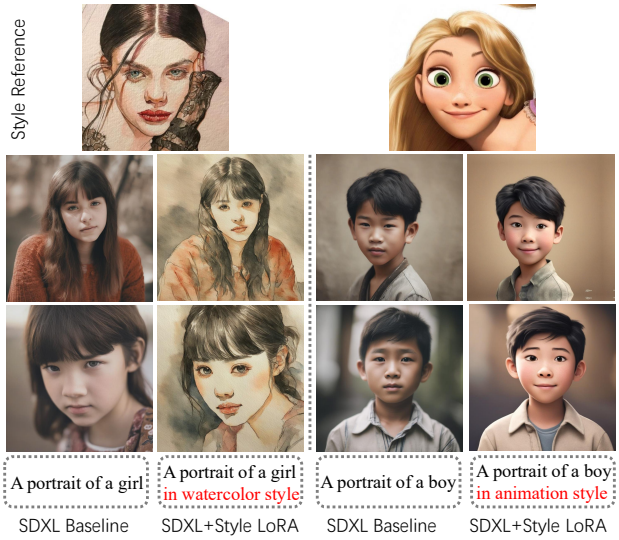

Figure 4. Leveraging the pre-trained SDXL backbone and SDXL backbone with trained style LoRA, the proposed method is capable of synthesizing paired photo-realistic image and stylized image by using the same random seed respectively.

4. Experiments

Implementation Details. In this paper, we use SDXL [22] as the backbone diffusion model, and LoRA weights for all styles are trained with AdamW optimizer [18] at learning rate 1×10^{-5} and batch size=1. We first train the geometry LoRA on the content images for 250 steps, and then jointly train geometry LoRA and style LoRA for another 250 steps. Training takes $\sim$10 minutes on a RTX6000 GPU. During inference, we employ a VLM [35] to extract detailed description of input images as text prompts for the diffusion backbone, and integrate ControlNet with canny edge into the framework for image-to-image translation.

Dataset and Pre-processing. We collect style examples from the internet and use FFHQ [15] dataset as content images, with image number 0-999 for test and image number 1000-9999 for train. All images are firstly resized to 544*544 resolution, after alignment and TPS warping, they are center-cropped to 512*512 size to avoid the artifacts near the edge area caused by warping. 28 facial landmarks are used for image alignment, 8 edge points of the rectangle images are combined with landmarks for warping.

Baselines. We compare our proposed method with JoJoGAN [3], InstantStyle [30], InstantID, OSASIS [2], B-LoRA [4], ZePo [17], and K-LoRA [20]. All the compared methods are tested for one-shot stylization. The comparison is shown in Figure 6. We use the official implementation for each method and test with released weights or train with de-

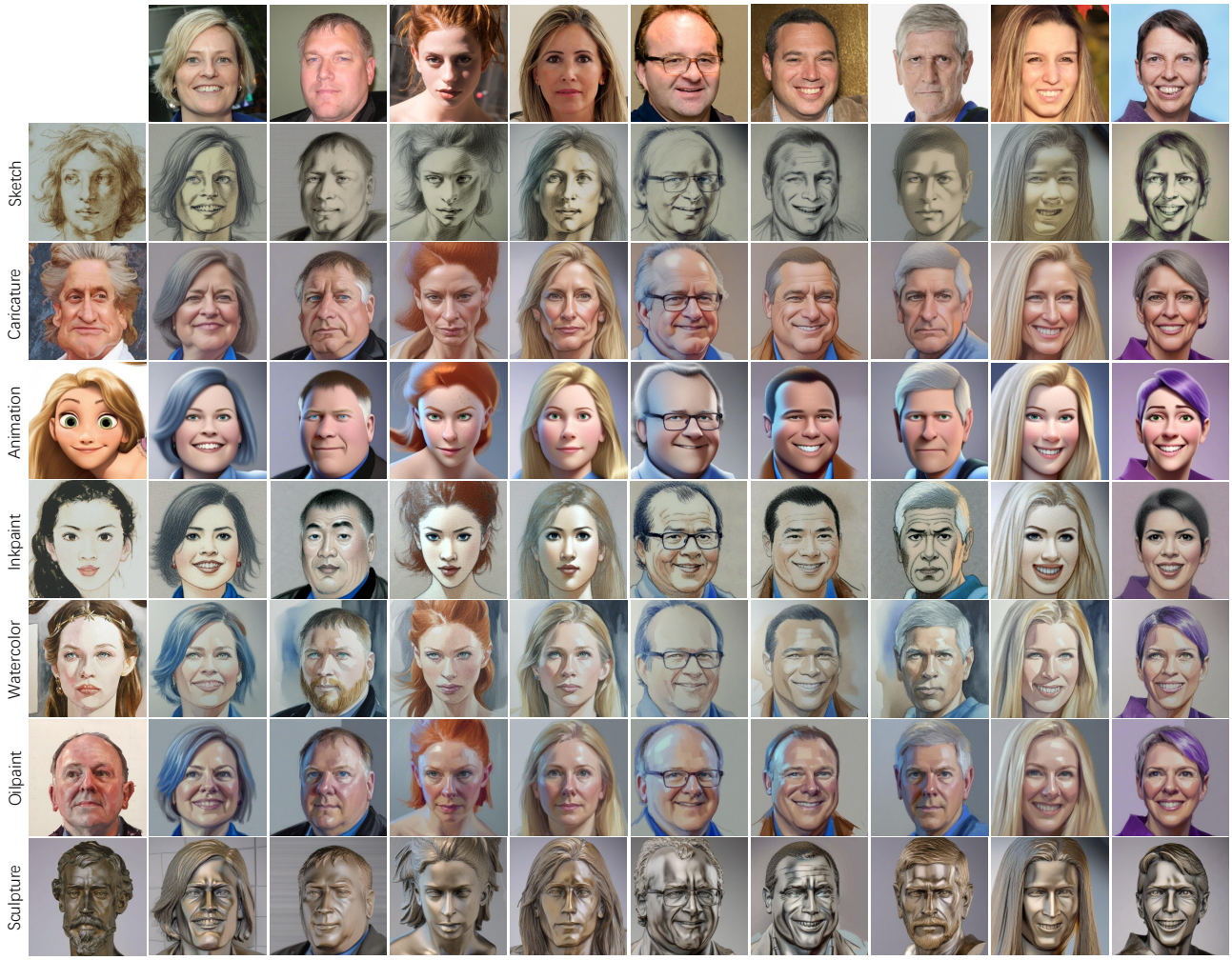

Figure 5. The proposed method synthesize high quality stylized portrait images across various styles.

fault settings and the same dataset used for our method.

Evaluation Metrics.

We employ Art-FID [34] and VGG style loss [5] to evaluate the stylization quality, and ArcFace cosine similarity [23] and LPIPS [41] to evaluate the content preservation.

Illustration of Stylization Results in Various styles. We show the stylization results of the proposed method in sketch style, caricature style, animation style, inkpaint style, watercolor style, oilpaint style and sculpture style in Figure 5. Our method consistently generates high-quality outputs across these styles while effectively handling portraits with diverse facial features and attributes.

4.1. Qualitative Comparison

We present qualitative comparisons between our method and several existing approaches in Figure 6. **JoJoGAN** introduces artifacts in the eye regions of the animation style and omits key elements, such as the hat in the caricature style, highlighting the limitations in structural preservation inherent to StyleGAN inversion-based methods. **InstantStyle** exhibits weak stylization effects for animation style and severe artifacts for caricature style, suggesting limited effectiveness for styles with strong geometric deformation or non photo-realistic references. **InstantID** retains identities but alters facial position and geometry, which undermines its suitability for stylization tasks where structural fidelity is essential. **OSASIS** produces high-quality results for watercolor and oil painting, yet fails to effectively learn and replicate the characteristics of caricature and animation styles, indicating difficulty in handling non photo-realistic domains. **B-LoRA** synthesizes rich textures but struggles to preserve the structural integrity and cases geometric distortion, such as losing the hat in caricature style and failure in reconstructing human faces in watercolor style results shown in the supplementary material. **ZePo** synthesizes high quality stylized textures and colors, but fails to maintain the person's age in animation style and identity in watercolor style. **K-LoRA** generates distorted textures in animation style and caricature style, and fails to maintain the face pose of oilpaint style.

In contrast, **Our method** demonstrates clear superiority

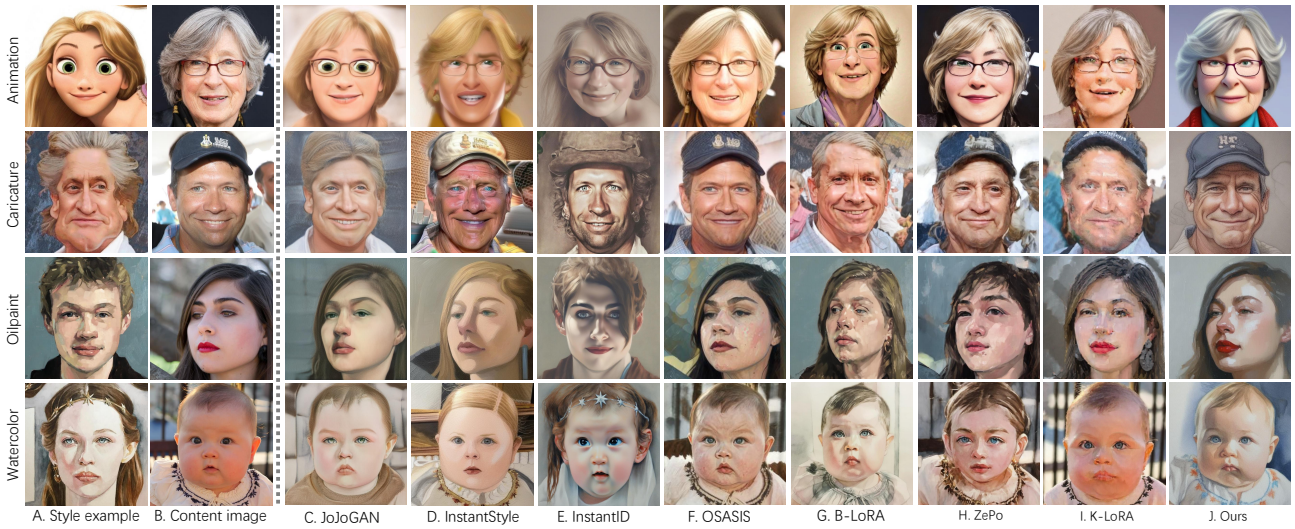

Figure 6. Qualitative comparisons of our method and existing methods. Our method synthesizes high quality results that represent clear and pleasant style and meanwhile loyally preserve the identity and structural information of content images.

Table 1. Quantitative comparison of the proposed model and existing methods. We use ArtFID↓ and VGG Style Loss↓ to evaluate stylization quality and ArcFace cosine similarity↑ and LPIPS Loss↓ to evaluate the content preservation compared to the content portrait images. red and blue represent the best and second best performance. The results are calculated across Animation, Caricature, Oilpaint and Watercolor styles. Our proposed method achieves the best performance in most of the comparison.

Style	JoJoGAN	InstantStyle	InstantID	OSASIS	B-LoRA	ZePo	K-LoRA	Ours
ArtFID↓	232.64	255.08	243.14	261.19	230.93	191.43	202.17	188.71
VGG Style Loss↓	4.6738	5.4716	5.1897	5.9371	4.9371	4.2736	4.5083	4.4619
ArcFace similarity↑	0.6877	0.6747	0.6392	0.7017	0.6828	0.6904	0.7005	0.7293
LPIPS↓	0.5895	0.6169	0.6907	0.4778	0.7125	0.4841	0.4996	0.4652

over all existing baselines by synthesizing high-quality stylized results that not only convey the intended style through vivid colors and textures, but also faithfully preserve the identity and geometric structure of the original content images. This performance can be attributed to the incorporation of geometric alignment, which establishes precise spatial correspondence between each content-style pair. This alignment facilitates the orthogonal adaptation process, enabling effective disentanglement of content and style representations. As a result, the model can jointly optimize separate LoRA weights, allowing it to learn style-specific characteristics while keeping the content information intact.

4.2. Quantitative Evaluation

Art-FID measures the distance of distributions in style feature space of a network pre-trained on art dataset, while VGG loss calculates the style gaps between the results and style references in style transfer tasks. ArcFace similarity compute the cosine similarity of image embeddings extracted by ArcFace network of image pairs, and LPIPS accesses the spatial similarity of images in the feature spaces of pre-trained image recognition network. In this work, we use Art-FID and VGG style loss to assess stylization quality and ArcFace cosine similarity and LPIPS loss to evaluate the preservation of structural and identity-related information. Lower Art-FID, VGG loss, LPIPS, and higher ArcFace similarities indicate better performance.

We perform quantitative comparisons across four quantitative creteria on caricature style, and summarize the results in Table 1. Our method achieves the best performances on ArtFID, ArcFace similarity and LPIPS and ranks the second on VGG style loss. These results quantitatively demonstrate that our approach effectively renders vivid and faithful stylization while preserving the structural and identity of the original portraits.

4.3. Ablation Study

Ablation study on training techniques are shown in Figure 7. Ablating orthogonality adaption causes severe artifacts and color shift in column A. This validates the importance of disentangling content information and style information, as the trainable LoRA weights fail to learn any meaningful information without orthogonality adaption. Ablating geometric alignment causes distorted month of the girl and color artifacts on the man's face in column B, using one-way geometric alignment to warp content images to style landmarks reduces artifacts in column C, but the qualities are still inferior to the full model in column D, which features distinct and pleasant animation style, bright color, fine textures, and free of artifacts. The comparison

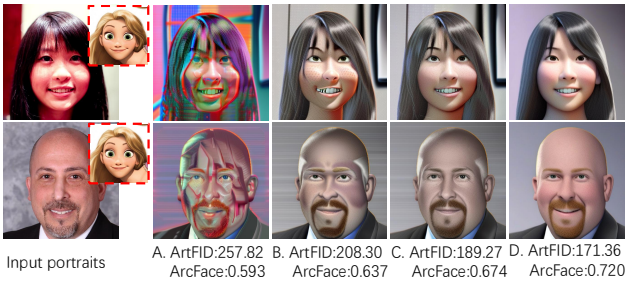

Figure 7. Ablation study of training techniques. We show qualitative results and qualitative results of ArtFID and ArcFace similarity of A: Ablating geometric alignment; B: one-side geometric alignment; C: Ablating orthogonality adaption; and D: full model.

Figure 8. Ablation study of baseline and prompt. We show qualitative results and qualitative results of ArtFID and ArcFace similarity of A: SDXL backbone; B: Ablating style prompt; C: Simple content prompt and D: Detailed content prompt.

of column B, C and D validates the effectiveness of geometric alignment: the geometric discrepancies between content-style pairs hinder stylization. The stylization effects improves when content-style pairs become more similar in geometry, as the LoRA effectively learns style information from aligned corresponding regions.

We also show the ablation study of SDXL baseline and prompts in Fig 8. The SDXL fails to synthesize results with proper style in A, indicating the stylization ability does not come from the backbone but the trained LoRA weights. Ablating style prompt causes stripe artifacts and wired color in B, showing that the LoRA weights need to work with style prompts. Using simple content prompt ("a portrait of a person") results in suboptimal results such as gender obfuscation in C, while detailed content prompt ("a portrait of a mid-aged white women with blonde hair") leads to visually pleasant stylization results with structure and identity unchanged, validating the combination of VLM extracted detailed prompts and ControlNet is compelling in preserving structure and identity information of the content image.

4.4. User Study

We utilize an user study to demonstrate how people evaluate the proposed method and existing methods. 40 participants are invited to select the best results with two criteria: the overall stylization quality and the preservation of identities information. We prepare 48 images with 12 images for each style, and randomly choose 8 images in each style to show a candidate. For each image, participants are presented the stylization result of the proposed method together with results generated by six existing methods.

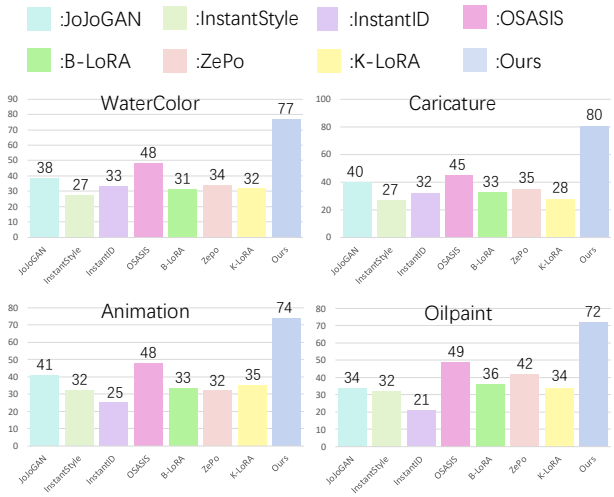

Figure 9. Results of user study. Our method is preferred across all methods in stylizaiton quality and identity preservation.

We present the results of the user study in Figure 9. Our proposed method has received the highest number of preferences compared to all the baseline illustrated. To further support the comparison, we employ the Kruskal-Wallis test for statistical test. The results clearly demonstrate that the proposed method significantly outperforms all existing methods in terms of user preference with a significance level of $p < 0.05$. The statistic results clearly show that our method receives distinctive preference by users investigated. All images shown in the user study are presented in the supplementary materials.

5. Conclusion

We propose a portrait stylization method that learns style information from a single style reference. We employ geometric alignment to align content images and the style reference to the mean landmarks and orthogonality adaption to disentangle the geometry and style information. A geometry LoRA and a style LoRA are jointly optimized with the aligned content-style pairs. During inference, we use style guidance to replace the standard classifier-free guidance and combine the diffusion model with ControlNet for image-to-image translation. The model can be trained for only a few hundred steps and synthesize high-quality stylized portraits with a single style reference. Qualitative comparison, quantitative evaluation and user study demonstrate that our proposed method outperforms existing methods. Ablation study shows the effectiveness of each component.

References

[1] Jia-Ren Chang and Yong-Sheng Chen. Exploiting spatial relation for reducing distortion in style transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1209–1217, 2021. 2

[2] Hansam Cho, Jonghyun Lee, Seunggyu Chang, and Yonghyun Jeong. One-shot structure-aware stylized image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8302–8311, 2024. 2, 3, 5

[3] Min Jin Chong and David Forsyth. Jojogan: One shot face stylization. In *European Conference on Computer Vision*, pages 128–152. Springer, 2022. 2, 3, 5

[4] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, pages 181–198. Springer, 2025. 3, 5

[5] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 2, 6

[6] Suryabhan Singh Hada and Miguel A Carreira-Perpinán. Style transfer by rigid alignment in neural net feature space. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2576–2585, 2021. 2

[7] Fangzhou Han, Shuquan Ye, Mingming He, Menglei Chai, and Jing Liao. Exemplar-based 3d portrait stylization. *IEEE Transactions on Visualization and Computer Graphics*, 29(2):1371–1383, 2021. 3

[8] Aaron Hertzmann. Painterly rendering with curved brush strokes of multiple sizes. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 453–460, 1998. 2

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 2

[14] Maxwell Jones, Sheng-Yu Wang, Nupur Kumari, David Bau, and Jun-Yan Zhu. Customizing text-to-image models with a single image pair. *arXiv preprint arXiv:2405.01536*, 2024. 2, 3

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 5

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[17] Jin Liu, Huaibo Huang, Jie Cao, and Ran He. Zepo: Zero-shot portrait stylization with faster sampling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3509–3518, 2024. 5

[18] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[19] Cewu Lu, Li Xu, and Jiaya Jia. Combining sketch and tone for pencil drawing production. In *Proceedings of the Symposium on Non-Photorealistic Animation and Rendering*, pages 65–73. Eurographics Association, 2012. 2

[20] Ziheng Ouyang, Zhen Li, and Qibin Hou. K-lora: Unlocking training-free fusion of any subject and style loras. *arXiv preprint arXiv:2502.18461*, 2025. 5

[21] Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7964–7973, 2024. 2, 4

[22] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3, 5

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[25] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[26] YiChang Shih, Sylvain Paris, Connelly Barnes, William T Freeman, and Frédo Durand. Style transfer for headshot portraits. 2014. 3

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[28] Aneta Texler, Ondřej Texler, Michal Kučera, Menglei Chai, and Daniel Sỳkora. Faceblit: Instant real-time example-based style transfer to facial videos. *Proceedings of the ACM*

on Computer Graphics and Interactive Techniques, 4(1):1–17, 2021. 3

[29] Bin Wang, Wenping Wang, Huaiping Yang, and Jiaguang Sun. Efficient example-based painting and synthesis of 2d directional texture. *IEEE Transactions on Visualization and Computer Graphics*, 10(3):266–277, 2004. 2

[30] Haofan Wang, Matteo Spinelli, 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, 3, 5

[31] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 3

[32] Xinrui Wang, Zhuoru Li, Xuanyu Yin, Xiao Zhou, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. Real-time data-efficient portrait stylization via geometric alignment. *Neural Networks*, page 107774, 2025. 2

[33] Xinrui Wang and Jinze Yu. Learning to cartoonize using white-box cartoon representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8090–8099, 2020. 2

[34] Matthias Wright and Björn Ommer. Artfid: Quantitative evaluation of neural style transfer. In *DAGM German Conference on Pattern Recognition*, pages 560–576. Springer, 2022. 6

[35] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 5

[36] Li Xu, Cewu Lu, Yi Xu, and Jiaya Jia. Image smoothing via l 0 gradient minimization. In *ACM Transactions on Graphics (TOG)*, volume 30, page 174. ACM, 2011. 2

[37] Dingkun Yan, Xinrui Wang, Zhuoru Li, Suguru Saito, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. Enhancing reference-based sketch colorization via separating reference representations. *arXiv preprint arXiv:2508.17620*, 2025. 2

[38] Dingkun Yan, Xinrui Wang, Zhuoru Li, Suguru Saito, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. Image referenced sketch colorization based on animation creation workflow. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23391–23400, 2025. 2

[39] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 5

[40] Lvmin Zhang, Xinrui Wang, Qingnan Fan, Yi Ji, and Chunping Liu. Generating manga from illustrations via mimicking manga creation workflow. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5642–5651, 2021. 2

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of IEEE International Conference on Computer Vision*, 2017. 2