1-2024

Learning an interpretable stylized subspace for 3D-aware animatable artforms

Chenxi ZHENG

Bangzhen LIU

Xuemiao XU

Huaidong ZHANG

Shengfeng HE
Singapore Management University, shengfenghe@smu.edu.sg

Learning an Interpretable Stylized Subspace for 3D-aware Animatable Artforms

Chenxi Zheng, Bangzhen Liu, Xuemiao Xu, Huaidong Zhang, and Shengfeng He, *Senior Member, IEEE*

Abstract—Throughout history, static paintings have captivated viewers within display frames, yet the possibility of making these masterpieces vividly interactive remains intriguing. This research paper introduces 3DArtmator, a novel approach that aims to represent artforms in a highly interpretable stylized space, enabling 3D-aware animatable reconstruction and editing. Our rationale is to transfer the interpretability and 3D controllability of the latent space in a 3D-aware GAN to a stylized sub-space of a customized GAN, revitalizing the original artforms. To this end, the proposed two-stage optimization framework of 3DArtmator begins with discovering an anchor in the original latent space that accurately mimics the pose and content of a given art painting. This anchor serves as a reliable indicator of the original latent space local structure, therefore sharing the same editable predefined expression vectors. In the second stage, we train a customized 3D-aware GAN specific to the input artform, while enforcing the preservation of the original latent local structure through a meticulous style-directional difference loss. This approach ensures the creation of a stylized sub-space that remains interpretable and retains 3D control. The effectiveness and versatility of 3DArtmator are validated through extensive experiments across a diverse range of art styles. With the ability to generate 3D reconstruction and editing for artforms while maintaining interpretability, 3DArtmator opens up new possibilities for artistic exploration and engagement.

Index Terms—3D-aware GANs, stylized animation, facial attribute editing.

1 Introduction

Artistic works serve as a reflection of artists' inspirations and perspectives on life, expressed in a creative and imaginative manner that has evolved alongside human history. While these artworks encapsulate profound introspections, subjective interpretations of existence, and fervent admiration for life, the stories they tell remain unheard. As a result, traditional static artworks often fail to resonate with ordinary viewers, hindering their ability to appreciate the inner spirit and beauty encapsulated within. However, in the era of deep learning, advancements in technology have made it more accessible for regular users to customize and stylize artistic content. This transformative process is gradually reshaping the way people create, consume, and share art, from everyday applications such as style image manipulation to virtual concepts like the metaverse. Consequently, there is a growing demand to rebuild the interactions between artists and audiences. This leads us to an intriguing question: "Can we breathe life into existing artworks, facilitating more vivid and resonating human-art interactions, all while preserving their original content and style?" This question gives rise to the problem of "*3D Animatable Artforms*", which involves the extraction of three-dimensional facial information from the given artistic face images while simultaneously maintaining the integrity of their content and style.

The unique properties of artistic portraits, however, introduce significant challenges in the realm of animating 3D artforms. Artistic representations, often rendered from a side view, frequently involve severe facial occlusion. This aspect makes it particularly challenging to accurately recover details in the parts of the face that are not visible. Moreover, the task of collecting multi-view consistent portraits is rendered impractical due to the inherent uniqueness of artistic works. In addition, the distinctive and personalized artistic styles, characterized by elements such as brush strokes, outlined edges, and exaggerated forms, create a substantial domain gap when juxtaposed with real human faces, further complicating the problem-solving process in this context.

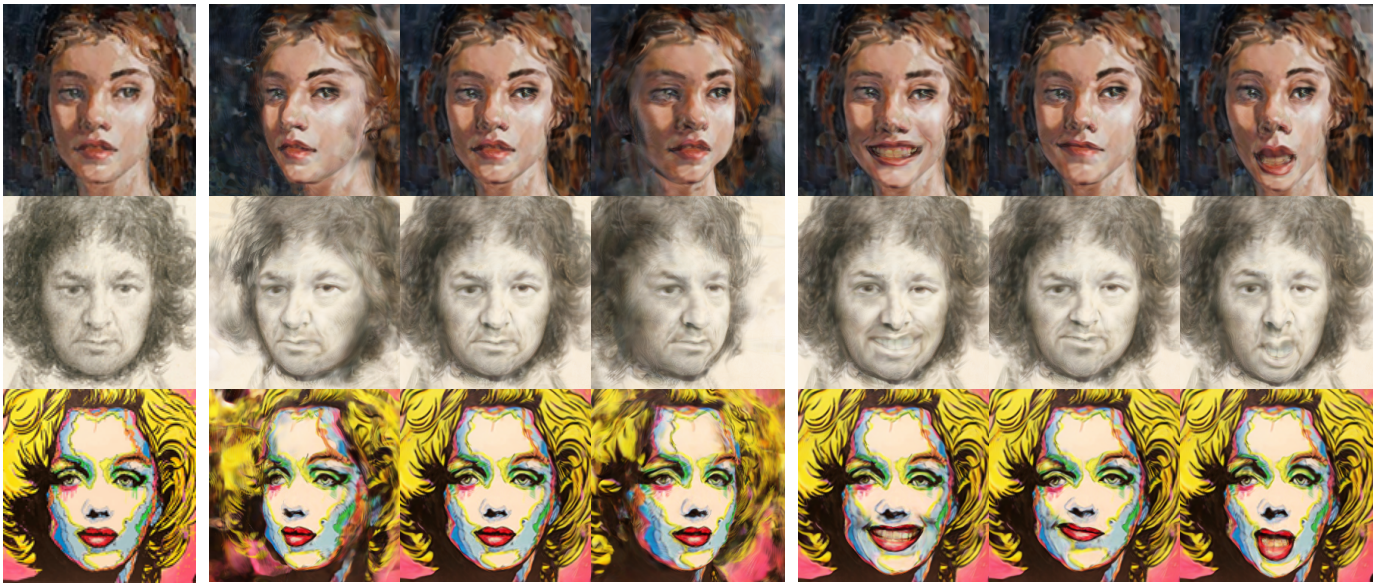
Neural stylization [15], [19], [34], [38] is an effective technique for emulating styles from diverse images [4], [12], [24], [25] or specific domains [7], [34], [38] onto target contents. The ability of deep networks [19], [25] to capture and change visual styles has been demonstrated in 2D image stylization, such as creating a picture in the style of Van Gogh's iconic starry sky. However, the application of these methods in 3D stylization under unrestricted camera poses is hindered by their limitation in stylizing the given view of the content image. Recent advances in 3D content stylization [16], [31], [61] have focused on leveraging the superior capabilities of neural radiance field (NeRF) [28]. These approaches have yielded impressive outcomes, characterized by enhanced visual quality and consistent geometric detail. To address the

The work is supported by the Guangdong International Technology Cooperation Project(No.2022A0505050009), China National Key R&D Program (No. 2023YFE0202700), Key-Area Research and Development Program of Guangzhou City (No.2023B01J0022), Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2023B1515020097), Singapore MOE Tier 1 Funds (MSS23C002), and the National Research Foundation Singapore under the AI Singapore Programme (No: AISG3-GV-2023-011). (Chenxi Zheng and Bangzhen Liu contributed equally.) (Bangzhen Liu and Xuemiao Xu are the corresponding authors.)

Chenxi Zheng, Bangzhen Liu, and Xuemiao Xu are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. E-mail: liubz.scut@gmail.com, cszcx@mail.scut.edu.cn, and xuemx@scut.edu.cn.

Huaidong Zhang is with the School of Future Technology, South China University of Technology, Guangzhou, China. E-mail: huaidongz@scut.edu.cn.

Shengfeng He is with the School of Computing and Information Systems, Singapore Management University, Singapore. E-mail: shengfenghe@smu.edu.sg.



(a) Input (b) 3D Reconstruction (c) Expression Editing

Fig. 1: We propose 3DArtmator that learns an interpretable stylized sub-space to enable 3D-aware reconstruction and editing of portraits in arbitrary artforms.

demand for generating 3D content, additional research has delved into the stylization of 3D GANs, such as EG3D [2] and GRAM [8]. Nevertheless, prevailing methods mainly focus on encoding the style of reference images while disregarding the remaining content, leading to style-similar but content-irrelevant images that fail to align with our requirements.

A straightforward solution for animating a given artwork is to combine GAN inversion [39], [49], [52] and cross-domain translation [11], [22], [66]. A previous method [18] has adopted this idea by training the entire generative model to predefined artistic styles. Content-preserving novel views are synthesized by inverting a reference image to its most similar latent code. However, the adversarial training involved in GAN's domain adaptation necessitates numerous images with the same style. Given the diverse categories of art styles worldwide, integrating all styles into a single model becomes an extremely challenging task. Consequently, the model must be adapted separately for each style, leading to the impractical and costly issue of data collection.

To achieve a more efficient and user-friendly 3D art animation, we propose reducing the training cost by animating artwork with just a single input image. Optimizing a latent code to simultaneously reconstruct invisible content and produce images consistent with a specific art style is impractical, primarily due to the substantial domain gap between the realms of real facial features and artistic styles. On the other hand, even though the original latent space might produce visually similar results to the given art, they suffer from an out-of-domain problem [11], [47], [48] that limits their interpretability and editability. This limitation restricts their applications in exploring more practical artistic possibilities.

In this paper, we propose 3DArtmator, an approach to establish an interpretable stylized space of a generator that exhibits the same editability as a 3D-aware GAN in the real domain. Our idea is to mimic the original latent space structure within a new stylized space. However, achieving a perfect alignment between the stylized space and the original latent space with a single style image is not possible. Instead of aiming for global consistency, we argue that achieving a locally interpretable sub-region suffices our one-shot setting. Consequently, our 3DArtmator adopts a two-step, two-generator training architecture. In the first step, we locate a local anchor and its corresponding structure that minimizes the transfer cost between the two spaces. In the second stage, we enforce the one-shot training of a customized 3D-aware GAN, which shares the same stylized subspace structure as the local sub-region, guided mainly by a style-directional difference loss. This objective function aims to maintain co-linearity with the directions between the different spaces. Our proposed 3DArtmator enables easy reconstruction of a diverse range of artforms in 3D while preserving interpretability within the stylized subspace. This leads to flexible and coherent 3D artistic editing without the need for a massive amount of training data.

2 Related Work

Generative 3D-aware Image Synthesis. Neural Radiance Fields (NeRF) have been introduced into 3D-aware Generative Adversarial Networks (3D GAN) [2], [3], [8], [40], showcasing impressive capabilities in high-fidelity novel view synthesis. Pioneering studies such as GRAF [40] and pi-GAN [3] aim to generate radiance fields from unstructured 2D images for 3D-aware image synthesis through differentiable volumetric rendering. Subsequent works have further improved synthesis quality by employing two-stage rendering [2], [13], [33] or exploring novel 3D representations [8], [41], [55], [65]. Despite significant advancements

in 3D GANs, direct application to artistic images fails to recover the underlying 3D geometry due to the substantial domain gap between real photographs and intricate artistic paintings.

Neural Stylization Deep neural stylization architectures have achieved remarkable results in transferring styles from either arbitrary images [4], [12], [15], [19], [24], [25], [43] or predefined domains [7], [34], [38], [56]. By incorporating temporal consistency with defined optical flows [6] or long-term correlations [1], [5], [42], neural style transfer techniques have also been extended to videos. However, existing image and video stylization methods are limited in their consideration of specific views. Without proper integration of 3D geometry representations, these neural stylization methods often suffer from heavy blurriness and cross-view inconsistencies when applied to off-the-shelf novel view synthesis frameworks. To achieve more robust and multi-view consistent stylization, recent works have employed Neural Radiance Fields (NeRF) [28] as implicit scene descriptors. By designing stylization networks, these approaches enable NeRF to predict color-related parameters based on a given reference image [10], [16], [31], [61] or textual descriptions [51] of the desired style. However, learning the stylization network typically requires multiple views of the same scene as input constraints. In contrast, our proposed approach enables the animation of any artforms using only a single image.

Domain Adaptation on GANs. Transferring GANs across domains has proven to be practical in both 2D [11], [17], [50], [59] and 3D-aware [14], [18] image synthesis through domain adaptation. In this process, a GAN model pretrained on real images is further fine-tuned on the target dataset, resulting in excellent results in synthesizing diverse and high-fidelity images with different styles. With the inheritance of the semantic-meaningful latent space, the adapted model also preserves the editing ability of the original GANs. However, existing adaptation techniques often require large-scale datasets for adversarial training, which can be time-consuming and impractical for certain artforms that may require professional guidance. Recent studies on few-shot [22], [29], [32] and one-shot domain adaptation [64], [66] have attempted to modify the GAN's style using only a small number of images or even a single image. Instead of adapting the entire GAN space to the target style, we aim to explore the generative ability and bring arbitrary artistic images to life while maintaining consistency across views and preserving content.

3 METHOD

3.1 Overview

For a given artform, the objective of our method is to learn a 3D interpretable representation of the stylized portrait. Since most of the artforms are delivered in one shot, the lack of 3D information raises a challenge to rebuild the stylized portrait from novel views. To learn the 3D representation of the stylized portrait solely from single-view supervision, our insight is to leverage the 3D priors embedded within 3D-aware GANs [2], [8], [55]. These 3D priors provide an interpretable latent space pretrained on the real portrait domain [21]. By incorporating these 3D priors and leveraging the interpretability of pretrained 3D-aware latent space, we gain the capability to effectively rebuild and manipulate portraits, thereby enhancing the fidelity and editability of the resulting 3D portraits.

Drawing upon the aforementioned insight, our proposed 3DArtmator aims to acquire an interpretable stylized space, facilitating the editing of a given painting in a manner that ensures multi-view consistency and style transferability. We propose to mimic a locally interpretable sub-region of the original latent space within the target style space. The optimization involves two progressive stages, including *Content-Analogous Subspace Localization (Sec. 3.2) and Directional-Guided Subspace Stylization (Sec. 3.3).* In the first stage, we recover the 3D information of the reference artwork via a content-analogous identification loss. This is achieved by optimizing a real portrait that closely resembles the content depicted in the given painting as an anchor in the 3D-aware GANs' local space. For the second stage, capitalizing on the multi-view consistency of the real domain, we introduce a style-directional difference loss, which ensures the co-linearity of the style directions across multiple views. Such that the local space can be effectively transferred from the real domain to the art domain while maintaining interpretability and consistent style characteristics. The overall architecture of our 3DArtmator is shown in Fig. 2.

3.2 Content-analogous Subspace Localization

Reconstructing a 3D-aware portrait presentation without multi-view supervision is challenging. To avoid the complex and deficient estimation of spatial properties (*e.g.*, 3D geometry and depth), we leverage the hidden 3D priors of the pretrained 3D-aware GANs and convert the estimation of 3D properties into a local sub-region location problem. Specifically, our goal is to optimize an anchor in the latent space of 3D-aware GANs, whose nearby latent are interpretable and have less cost on adapting to the stylized space of the referenced style. To achieve this, given a reference image of artform with a manually specified view ϑ as inputs, we search for the optimal latent code z^* in the real portrait space that produces the most identity-similar image to the reference image from the view ϑ.

Suppose we have a pretrained 3D generator G_{ori}, the process of rendering an image from a given view ϑ can be formulated as:

$$I_{ori,\vartheta} = G_{ori}(z, \vartheta). \tag{1}$$

Given a referenced portrait I_{ref}, we are searching for an anchor z that has less transferring cost to I_{ref}. The distance between I_{ref} and $I_{ori,\vartheta}$ are minimized in the perspective of style and structure, leading to the following content-analogous identification loss:

$$\mathcal{L}_{cr} = \lambda_1 \mathcal{L}_{perc}(I_{ori,\vartheta}, I_{ref}) + \lambda_2 \mathcal{L}_{pix}(I_{ori,\vartheta}, I_{ref}), \tag{2}$$

where $\mathcal{L}_{perc}$ and $\mathcal{L}_{pix}$ indicates the LPIPS loss [62] and L2-norm pixel-wise loss respectively. With the above restriction, the optimal anchor z^* in the pretrained latent space is localized by:

$$z^* = \arg\min_z \mathcal{L}_{cr}. \tag{3}$$

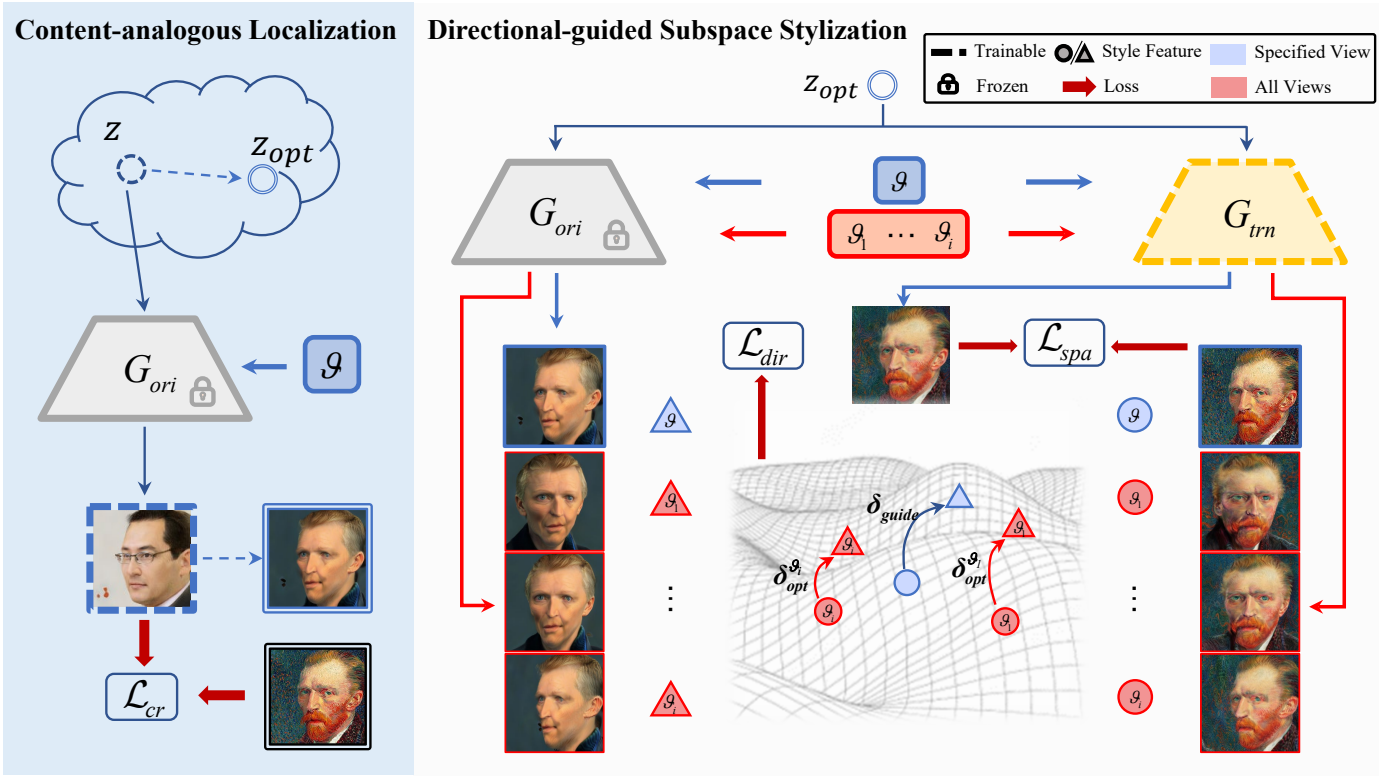


Fig. 2: Our proposed 3DArtmator consists of two progressive optimization steps, including the content-analogous subspace localization and the directional-guided subspace stylization. For localizing the content-analogous subspace, we optimize a random noise to generate the most identity-similar real portrait with the reference painting. For directional-guided subspace stylization, we adopt the style direction of the reference view between the real portrait and the reference painting as directional guidance. The other directions of generated views are forced to align with the guided style direction in the style space.

In this way, the optimal anchor collaborated with its subspace can preserve the content associated with the reference portrait. While maintaining the editability inherited from the original latent space, this subspace enables the subsequent transfer learning of the target stylized space.

3.3 Directional-guided Subspace Stylization

With the interpretable latent subspace localized by latent optimization in the real space, the following step is to transfer it to the stylized space of the given painting. Since we have only one artwork, directly learning the style in one shot will lead to severe overfitting and break the interpretability of the subspace. To address this concern, we propose to train the generator G_{trn} with the proposed style-directional difference loss.

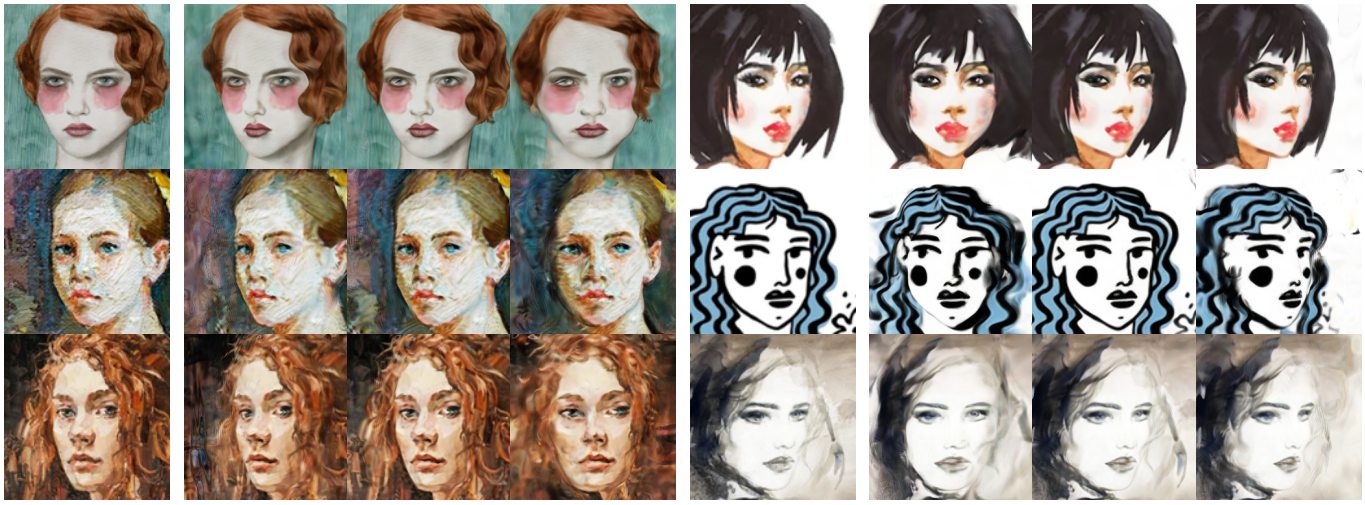
3.3.1 Style-directional Difference Loss

Compared with portraits from the real world, artworks contain more hand-crafted traces, introducing extra difficulty in embedding the style within the subspace. Directly applying a 2D spatial alignment (e.g., LPIPS loss) between the rendered target-style image at view ϑ and the portrait is helpful in 2D style transfer but fails in 3D scenarios. The one-shot supervision easily leads to deterioration in radiance rendering, such that the occluded radiances in the field are blurred or vitrified. Artifacts such as the sticker effect and tortuosity happen due to the absence of depth and limited contextual information, resulting in non-stereoscopic surfaces and asymmetric faces. Besides, an insufficient viewpoint will lead to the ambiguity of 3D property estimation, which deteriorates the interpretability of the original subspace.

To achieve stable style transfer for the subspace, we propose to boost the geometric consistency of the resulting stylized space with directional guidance and jointly optimize the 3D generator from multiple view constraints. Since the stylized portrait is available in only one view, we achieve a 3D-consistent artistic stylization optimization by extracting the style-directional difference from the stylized portrait. The style-directional difference is calculated between the anchor and the referenced portrait image in the stylized space and serves as a constraint for the stylization of arbitrary views. Previous text-guided translation methods guide the transfer by the difference between the feature of two style descriptions t_{src} and t_{tgt}, which is denoted as $\delta_t = E_T(t_{tgt}) - E_T(t_{src})$, where E_T represents the text encoder of a pretrained CLIP [36]. Inspired by this, we exploit the difference of the style features extracted from the image pair $I^*_{ori,\vartheta} = G_{ori}(z^*, \vartheta)$ and I_{ref} as directional guidance, which is denoted as:

$$\delta_{guide} = E(I_{ref}) - E(I^*_{ori,\vartheta}), \tag{4}$$

where E represents the feature encoder that projects the



(a) Input (b) Results (c) Input (d) Results

Fig. 3: 3D reconstruction with AniFaceGAN. (a) & (c) are the reference paintings, while (b) & (d) are the multi-view rendering results of the reconstructed 3D portrait. The images in the central column of (b) & (d) are the rendering results with the same view of the reference painting.

image into a style space $\mathcal{S}$. δ_{guide} is treated as a standard direction for optimization. We expect the directions of sample pairs from other unspecific views can maintain the co-linearity that is directionally aligned with δ_{guide}. To achieve this goal, we randomly sample N views $\{\vartheta_i\}_{i=0}^{N}$ and construct N image pairs $(I^*_{ori,\vartheta_i}, I^*_{trn,\vartheta_i})$ with the pretrained generator G_{ori} and the trainable generator G_{trn}, where

$$\begin{aligned} I^*_{ori,\vartheta_i} &= G_{ori}(z^*, \vartheta_i), \\ I^*_{trn,\vartheta_i} &= G_{trn}(z^*, \vartheta_i). \end{aligned} \tag{5}$$

The style directions are calculated for each views ϑ_i by:

$$\delta^{\vartheta_i}_{opt} = E(I^*_{trn,\vartheta_i}) - E(I^*_{ori,\vartheta_i}). \tag{6}$$

Our style-directional difference loss is calculated by the negative cosine similarity ($\langle\cdot,\cdot\rangle$) between the style guidance δ_{guide} and the set of $\{\delta^{\vartheta_i}_{opt}\}_{i=0}^{N}$ for optimization:

$$\mathcal{L}_{dir} = \sum_{i=0}^{N} 1 - \langle \delta_{guide}, \delta^{\vartheta_i}_{opt} \rangle. \tag{7}$$

3.3.2 Overall Objective

Noticed that the style-directional difference loss evaluates the co-linearity between two spaces but does not directly supervise the capturing of the spatial features of the art portrait. To produce a reconstructed 3D portrait that has consistent content with the reference image, we apply an extra spatial constraint on G_{trn}. Therefore, the spatial alignment loss concerning the latent code z^* and the specific view ϑ is defined as follows:

$$\mathcal{L}_{spa} = \mathcal{L}_{perc}(G_{trn}(z^*, \vartheta), I_{ref}), \tag{8}$$

where LPIPS is adopted as the perceptual loss.

Finally, we optimize the 3D generator G_{trn} with the following loss for cross-view stylization:

$$\mathcal{L}_{style} = \lambda_3 \mathcal{L}_{dir} + \lambda_4 \mathcal{L}_{spa}. \tag{9}$$

4 Experiment

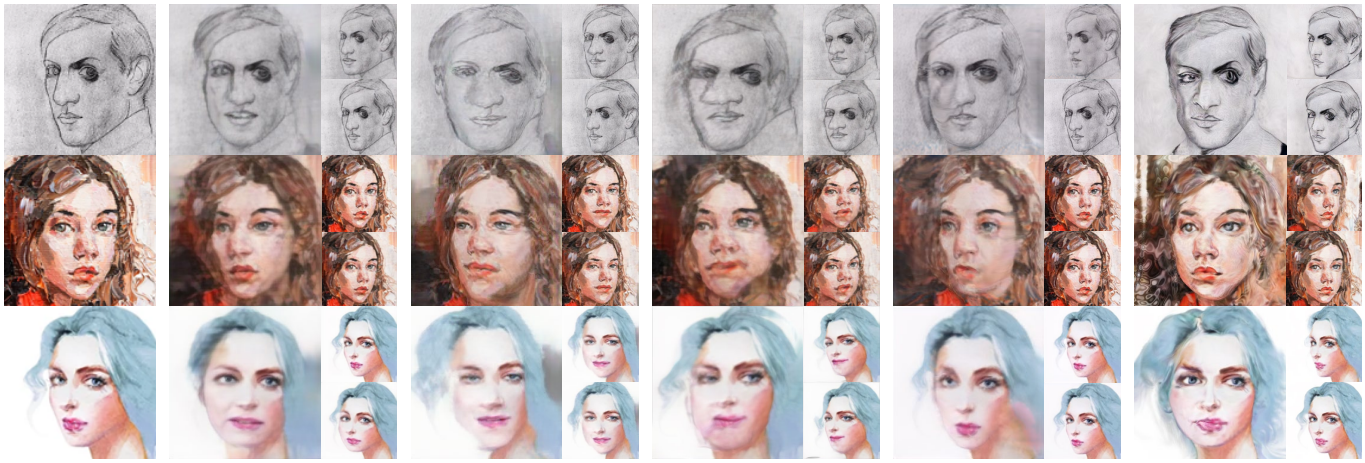
4.1 Implementation Details

We employ the CLIP [36] encoder to project the images into the stylized space. Here we adopt AniFaceGAN [55] as our backbone to facilitate its 3D prior and editatbility on the real domain. The rendered images I^*_{trn,ϑ_i} and I^*_{ori,ϑ_i} in Eq. (6) are both augmented before image encoding. In each training iteration, we randomly selected $N = 1$ in Eq. (7) for cross-view stylization. Moreover, we adopt Adam [23], [57], [58] as the optimizer in our 3DArtmator. In the first stage, the latent code z is initialized randomly and optimized with a learning rate of $2e-2$. In the second stage, the trainable generator G_{trn} is optimized with a learning rate of $1e-5$. The weight factors in Eq. (2) and Eq. (9) are all set as 1.0. The single optimization take is completed within 20 minutes using a single 40GB NVIDIA A100 Tensor Core GPU.

4.2 3D Reconstruction

In Fig. 3, we present several challenging scenarios encountered in the realm of 3D Artistic Animation. These instances share a common characteristic where the portraits deviate from the forward direction, resulting in significant occlusion. However, our 3DArtmator framework successfully achieves high-fidelity 3D reconstruction across a diverse range of artforms. When viewed from the specified angle ϑ, the rendered image exhibits remarkable visual similarities to the input, effectively capturing the distinctive artistic features, such as brush stroke styles. Notably, when observed from alternative perspectives, 3DArtmator excels in reconstructing identities, including the intricate estimation of facial attributes, such as the curvature of the jawline and the distinct stereo bridge of the nose. These outcomes highlight the efficacy of our approach in capturing and representing the intricate details of artistic portraits.

Despite the high fidelity, 3DArtmator demonstrates the ability to generalize to different artforms. Even when pro-



(a) Input (b) FOMM (c) PIR (d) MRAA (e) LIA (f) Ours

Fig. 4: Comparison of 3D reconstruction quality with face reenactment methods. The input artistic paintings are resized to 128×128 resolution. The rendering images with a large twisting angle related to the original view are presented on the left of each column, while the right two are with smaller view discrepancies.

vided with a single portrait as input, 3DArtmator successfully extracts the stylistic essence faithfully and efficiently guides the optimization process to align unspecified views with the artistic domain. This capability is exemplified in Fig. 3, where a diverse collection of artforms, including decorative painting, sketch, watercolor, and oil painting, showcases the remarkable generalizability of 3DArtmator across different painting styles.

4.3 Comparison

4.3.1 Comparison with Face Reenactment

We compare our method with existing face reenactment approaches, which were originally pretrained on VoxCeleb [30] and directly applied to artistic animation without further fine-tuning. To provide a wider range of observation views, the driving videos used for the reenactment approaches are generated using pretrained AniFaceGAN [55]. Specifically, we considered FOMM [44], PIRenderer [37], MRAA [45], and LIA [53] for comparison purposes. The visualization results are presented in Fig. 4.

Near the specified view ϑ, the reenactment methods exhibit the ability to produce high-quality images. However, as the viewpoint deviates further from the frontal orientation, identity alteration becomes evident in the reenactment results (Fig. 4b, Fig. 4c, and Fig. 4d), or geometric distortion occurs (Fig. 4e). These phenomena can be attributed to the inherent dependence on the geometric properties of the driving videos employed by FOMM, PIRenderer, and MRAA, which inevitably leads to a degradation in maintaining the reference identity. In contrast to the aforementioned reenactment approaches, our 3DArtmator method operates without driving videos and reconstructs a 3D-consistent portrait. As a result, our approach exhibits superior visual performance, particularly when faced with significant deviations in viewpoints.

We conducted a corresponding user study involving 20 video clips reconstructed by the above methods, with over 50 participants engaged in the study. They were tasked with ranking the visual quality of the videos based on three criteria: reconstruction fidelity, 3D consistency, and interpretability, respectively. As shown in Tab. 1, the statistical results are calculated by the average ranking of each method over all the participants. Overall, the user study confirms the superiority of our 3DArtmator method in terms of reconstruction fidelity and 3D consistency, while maintaining competitive performance in editing capabilities when compared to existing face reenactment approaches.

TABLE 1: A user study on fidelity, consistency, and interpretability. We report the average rank of each method with the standard deviation over all participants. The smaller values indicate the higher acceptance rate of the audiences. The best records are marked in **bold**.

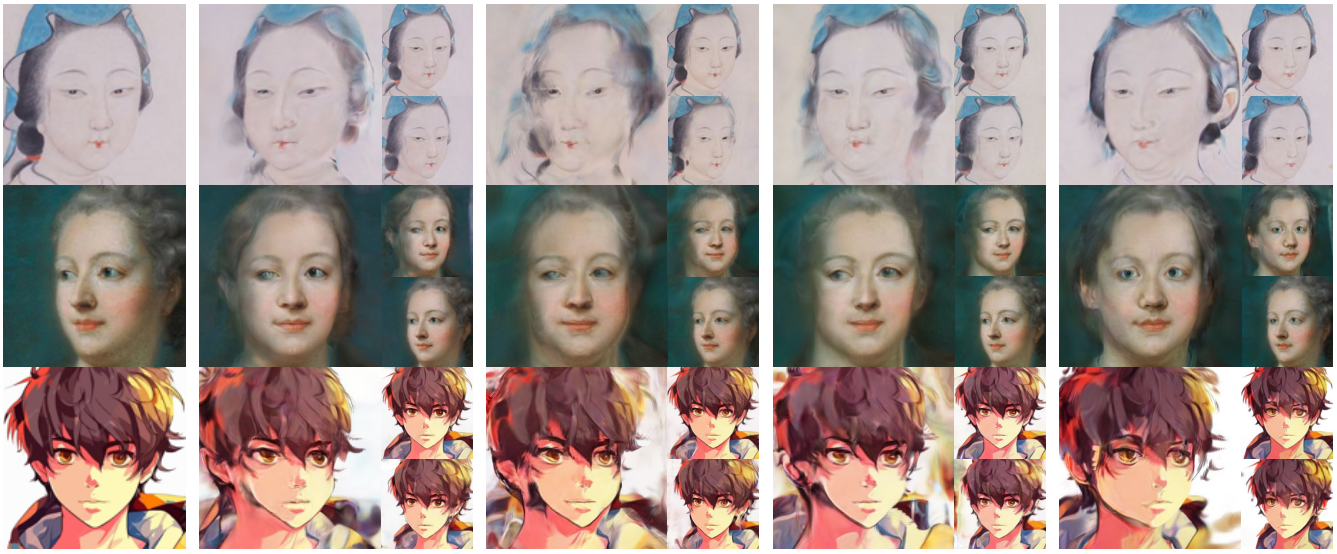
Methods	Fidelity	3D-Consistency	Interprebility
FOMM	2.61± 0.60	2.45± 0.65	2.23± 0.64
PIR	2.88± 0.31	3.05± 0.31	2.90± 0.31
MRAA	4.43± 0.32	4.36± 0.33	3.79± 0.32
LIA	2.60± 0.34	2.60± 0.32	2.57± 0.31
3DArtMator	**1.47± 0.31**	**1.65± 0.32**	**1.81± 0.31**

4.3.2 Comparison with AniFaceGAN-based Methods

To highlight the effectiveness of 3DArtmator, we conduct comparisons with other stylization methods across three distinct style domains: Ukiyoe [35], Metface [20], and Webtoon [1], as illustrated in Fig. 5. For each domain, we randomly select a set of 50 stylized images, denoted as Φ_{sty}, and apply two renowned stylization techniques: FreezeG [2] and StyleGAN-NADA [11], to the AniFaceGAN model. Subsequently, we utilize PTI [39] to optimize the latent code corresponding to the in-domain stylized portrait, enabling a comprehensive comparison of the stylization capabilities.

1. Naver Webtoon Datasets. https://github.com/bryandlee/naver-webtoon-data

2. Freezing generator for pseudo image translation. https://github.com/bryandlee/FreezeG.



(a) Input (b) PTI* (c) FreezeG+PTI* (d) NADA+PTI* (e) 3DArtmator

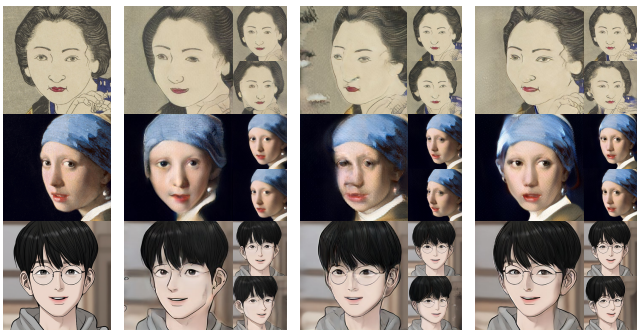
Fig. 5: Comparison with stylization methods based on AniFaceGAN. The styles from top to bottom are Ukiyoe, MetFace, and Webtoon, respectively. * denotes implemented by the authors.

For an in-depth analysis, we begin by evaluating the outcomes of directly applying PTI to the pre-trained AniFaceGAN. This contrasts with our subspace localization approach (detailed in Sec. 3.2). PTI incorporates a locality regularization term [39] that aims to limit significant changes in the latent space region surrounding the optimized anchor. However, the goal of stylization typically involves a comprehensive shift in the entire feature space. This makes the locality regularization term counterproductive, as it attempts to preserve the original feature space.

To better align PTI with the target style domain, we adjust the model by reducing the weight λ_{loc} of the locality loss to $1e-4$. Despite this adjustment, as shown in Fig. 5b, PTI still struggles with issues like pixel degradation and composition artifacts. 3DArtmator, on the other hand, effectively mitigates these problems through its directional guidance (more discussion can be found in Sec. 4.5).

As an alternative approach to mitigate artifacts, we explore pre-stylizing the pretrained AniFaceGAN before inversion. This involves adopting the adversarial transferring techniques of FreezeG and the CLIP guidance of StyleGAN-NADA, replacing their generators with AniFaceGAN in practice. In adversarial stylization, we freeze the first five out of eight SIREN [26] layers in AniFaceGAN and optimize the discriminator with the stylized image set Φ_{sty} from scratch using the non-saturating GAN loss [27]. For CLIP-guided stylization, we construct a set of real images Φ_{real}, matching the number in the stylized set Φ_{sty}. These real images are synthesized by AniFaceGAN to emulate the style of the real domain. The differences between Φ_{real} and Φ_{sty} are used to calculate the CLIP directions, which guide the optimization.

The results, showcased in Fig. 5c and Fig. 5d, reveal that while the stylized AniFaceGAN provides a strong style prior, it compromises reconstruction accuracy, even more than using PTI alone. This issue arises because stylization with limited data tends to reduce the expressiveness of the latent space, consequently degrading the quality of inversion. The less accurate anchor in such scenarios exposes subsequent pivotal tuning to significant feature discrepancies, leading to overfitting at optimized viewpoints. This observation is further supported by the quantitative evaluations in Sec. 4.4 and Sec. 4.3.4.



(a) Input (b) Dr. 3D (c) AgileGAN3D* (d) 3DArtmator

Fig. 6: 3D reconstruction of EG3D-based methods. The styles from top to bottom are Ukiyoe, MetFace, and Webtoon, respectively. (a) the reference input. (b) the results of Dr. 3D. (c) the results of AgileGAN3D (d) the results of our 3DArtmator. * denotes implemented by the authors.

4.3.3 *Comparison with EG3D-based Methods*

We compare 3DArtmator with Dr. 3D [18] and AgileGAN3D [46] in three artistic domains, *i.e.*, Ukiyoe [35], Metface [20], and Webtoon. Dr. 3D is a 3D-aware domain adaptation framework, which requires a large-scale set of portrait paintings from the same artistic domain. We directly utilize the pretrained checkpoints of Dr .3D and apply GAN inversion for portrait reconstruction. AgileGAN3D is an efficient two-stage few-shot stylization framework, including

TABLE 2: Quantitative comparison for 2D Face Reenactment methods (VIDEO DRIVEN) and AniFaceGAN-based methods (3DGAN). The best two records are marked in **bold** and underlined. * denotes implemented by the authors.

	Methods	LPIPS↓	PSNR↑	SSIM↑
VIDEO DRIVEN	FOMM	0.3695	20.8539	**0.6331**
	PIR	0.3718	21.0825	0.5838
	MRAA	0.4343	18.3001	0.4749
	LIA	0.3704	21.1195	0.6091
3DGAN	PTI*	0.3552	20.5301	0.6105
	FG+PTI*	0.3781	19.6122	0.5799
	NADA+PTI*	0.3579	20.0995	0.6078
	3DArtmator	**0.3428**	**21.1798**	0.6177

style prior creation and subsequent guided transfer learning. For the style prior creation, we first transfer the original StyleGAN2 with 20 stylized samples following StyleGAN-NADA [11] and construct a training set that contains 1,000 real-stylized sample pairs. This training set is utilized for transfer learning. During the guided transfer learning, we adopt a pretrained GOAE [60] encoder for 3D GAN inversion and finetune the pretrained EG3D with transfer learning loss and guidance loss. Note that for each new domain, both Dr. 3D and AgileGAN3D require extra training data for stylization, while 3DArtmator can efficiently render the 3D portrait from any single portrait image.

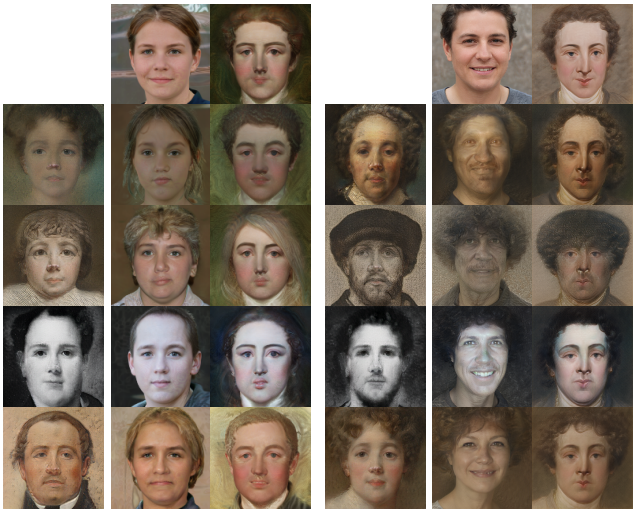
The qualitative results are shown in Fig. 6. Dr. 3D maintains a satisfying 3D consistency, but the identity of the reconstructed scene is not strictly the same as the input portrait. AgileGAN3D exhibits geometric artifacts in its reconstructions, a challenge akin to the issue of unfaithful inversion discussed in Sec. 4.3.2. 3DArtmator in Fig. 6d excels in achieving precise and 3D-consistent reconstruction. It reaches a level of reconstruction performance comparable to that of Dr. 3D, while not requiring an extensive collection of training data.

4.3.4 Quantitative Evaluation

Due to the lack of 3D stylized datasets, we construct a multi-views synthesis dataset with Dr. 3D to provide pseudo ground truth for quantitative evaluation. Specifically, N_q synthesized portraits $\{z^i | 0 \leq i \leq N_q\}$ are manually selected from a series of randomly sampled latent codes, which have no distinct artifacts and yield strict 3D consistency. For each portrait z^i, we further render images of N_v views with yaws ϑ^j uniformly sampled from $[\frac{\pi}{5}, -\frac{\pi}{5}]$. Therefore, the resulted image set is denoted as $\{x^{i,j} | 0 \leq i \leq N_q, 0 \leq j \leq N_v\}$, where $x^{i,j} = Q(z^i, \vartheta^j)$; Q and z^i represents the Dr. 3D generator and random noise, respectively. In practice, the values of N_q and N_v are set as 50 and 7, respectively.

To evaluate the performances of different models, we take the front view of each synthesized portrait $x^j = Q(z, \vartheta^j)$, where $\vartheta^j = 0$, as the input single-view images for 3D reconstruction. After obtaining the reconstructed portrait, we further render images for the rest $N_v - 1$ views for evaluation.

We denote the optimized latent code and generator for each scene as z^i and G^i. Following previous works [28], reconstruction accuracy s is calculated between the single-view generated image $\hat{x}^{i,j} = G^i(z^i, \vartheta^j)$ and pseudo ground-truth pairs with similarity metric $m(\cdot, \cdot)$, *i.e.*, $s = \mathbb{E}_{i,j}[m(\hat{x}^{i,j}, x^{i,j})]$. The evaluated metrics include LPIPS [62], PSNR, and SSIM [54]. Before calculating the pair-wise similarity, we crop the image and obtain the face region to eliminate the impact of background and implicit coordinate differences.



(a) AniFaceGAN (b) AgileGAN3D

Fig. 7: Results of GAN inversion to the latent space of pretrained / stylized 3D GANs. AniFaceGAN and AgileGAN3D are selected as examples. For each GAN, the leftmost column shows the input images, while the right part presents the inversion results to the pretrained latent space and stylized latent space, respectively. The top row presents the generated image of initial latent codes before inversion.

The quantitative results are presented in Tab. 2. We compare 3DArtmator with 2D face reenactment methods (top part in Tab. 2) and combinations of 2D stylization methods and AniFaceGAN (bottom part in Tab. 2). Following Sec. 4.3.1, we obtain the multi-view images with face reenactment methods by driving the stylized single input with a 3D consistent video. For 3DGAN experiments, all frameworks are implemented with the same backbone AniFaceGAN. Image stylization methods (FreezeG and StyleGAN-NADA [11]) are incorporated to transfer AniFaceGAN to the target domain before applying single image inversion [39]. Note that the extra stylized set Φ_{sty} introduced in Sec. 4.3.2 are provided during the stylization for FreezeG and StyleGAN-NADA, while PTI [39] and our 3DArtmator only rely on a single image.

4.4 Analysis of Latent Space Expressiveness

The reason for the powerful editability of GANs is the rich and comprehensive representation within the latent space. Thus, preserving the expressiveness of latent space during optimization is crucial for faithful and flexible 3D stylization. As previously discussed in Sec. 4.3.2 and Sec. 4.3.3, after transferring the generator to a given style, the generated results of existing stylization methods exhibit undesirable artifacts, such as pixel degradation and sticker effect. In this section, we conduct several experiments to

(a) AniFaceGAN (b) AgileGAN3D

Fig. 8: The diversity of latent space for the pretrained / stylized 3D GANs. For each GAN, we randomly sample 12 latent codes in the latent space, and present 12 generated images using the same latent codes in both the pretrained and stylized latent space. The results are listed in the left two columns and the right two columns, respectively. Note that the latent codes of the corresponding position in the left columns and the right columns are the same.

evaluate the latent spaces of these methods and 3DArtmator, and explain the undesirable behaviors mentioned above from the perspective of latent space expressiveness.

Considering AniFaceGAN and AgileGAN3D as examples, we first compare the inversion results of applying PTI to the pretrained latent space and the latent space after stylization, respectively. The results are shown in Fig. 7. We find that the images generated from the pretrained latent space exhibit more diverse details and suitable attributes. In contrast, images generated from the stylized latent space fail to preserve the identities of the input images, on the other hand, are much more similar to the images generated from the initial latent codes. These phenomenons indicate a potential degeneration of the pretrained latent space after stylization. The decline of its expressiveness results in the failure of the optimization-based GAN inversion approach.

We further investigate the original latent space of the pretrained 3D GANs G_{real} and the stylized one G_{sty} for more experimental evidence. We construct a set $\{(z^i, \vartheta^i)|0 \leq i \leq N_l\}$, which consisting of $N_l = 1,000$ pairs of randomly sample latent codes z^i and corresponding render views ϑ^i. Subsequently, we generate two paired image sets $\Phi_{real} = \{x^i_{real}|0 \leq i \leq N_l\}$ and $\Phi_{sty} = \{x^i_{sty}|0 \leq i \leq N_l\}$, where $x^i_{real} = G_{ori}(z^i, \vartheta^i)$ and $x^i_{sty} = G_{sty}(z^i, \vartheta^i)$. Qualitatively, we select 12 pairs of images from Φ_{real} and Φ_{sty} for visualization. As shown in Fig. 8, compared with the origin latent space, the images generated from the stylized latent space tended to be homologous, and exhibit similar appearances and textures. We also quantitatively evaluate the diversity of Φ_{real} and Φ_{sty} using the LPIPS. Concretely, d_{real} and d_{sty} are calculated by $d_{real} = \mathbb{E}_{i,j,i\neq j}[m(x^i_{real}, x^j_{real})]$ and $d_{sty} = \mathbb{E}_{i,j,i\neq j}[m(x^i_{sty}, x^j_{sty})]$, respectively, where $m(\cdot,\cdot)$ represents the LPIPS. The values of d_{real} and d_{gsty} are 0.5301 and 0.3902 for AniFaceGAN. For AgileGAN3D, d_{real} and d_{sty} are 0.6408 and 0.4790. The diversity of the latent space is decreased after stylization.

In conclusion, stylizing GANs without enough data deteriorates the expressiveness of latent space, leading to difficulties in localizing the latent code to the correct subspace. Therefore, the inversion cannot accurately reproduce the original structure details of the input image, resulting in geometric artifacts. Generally, it is hard to collect enough data on a certain style domain to transfer the latent space without degeneration. Compared with the aforementioned stylization methods, it is reasonable for 3DArtmator to first localize the subspace within the original latent space that has more expressiveness for better reconstruction, which also benefits the following stylization.

4.5 Ablation Study

We conduct ablation studies to verify the effectiveness of the proposed components, including the content-analogous subspace localization, style-directional guidance, and spatial constraint. We visualize three cases that have various complexities as shown in Fig. 9. *Mona Lisa* (row 1 in Fig. 9) is relatively simple as it possesses a facial structure similar to the real domain. *Red Madras Headdress* (row 2 in Fig. 9) is a fusion of vibrant colors and intricate patterns. The crux of this case is the feasibility of modeling the heavy outlines of edges in the 3D portrait. *Van Gogh's portrait* (row 3 in Fig. 9) is another type of hard case, as the facial structure and texture are distinct from the real domain. Therefore, it is imperative to evaluate the reconstruction of occluded parts.

There exists an enormous feature gap between the reconstructed portrait and the reference artistic image. Without the subspace localization, *i.e.*, randomly initializing the latent code and fixing it when stylizing the subspace, both the content and style need to be jointly optimized, leading to a complex optimization task of local facial attribute reconstruction. In the case of *Mona Lisa*, the 3D identity is initialized with a male, making the generator difficult to "grow the long hair" during the later stylization. Compared with the outcomes of the full model, this variant produces more geometrical artifacts, as highlighted in the dash areas in Fig. 9.

Without style-directional guidance, the reconstruction is supervised from only one perspective, and therefore the spatial information cannot be estimated correctly, resulting in geometrical artifacts. In terms of *Mona Lisa*, the obscured half of the nose cannot be elongated together with the visible half, leading to the asymmetry of facial attributes. *Red Madras Headdress* demonstrates a manifest sticker effect, where the rough stroke edges are not recognized as visual boundaries, but are directly attached to the skin as the texture. This impairs the stereoscopic effect and reduces vividness. In the case of *Van Gogh's self-portrait*, the occluded part suffers from severe pixel degradation. The unsupervised right face gradually degrades into transparent voxels, which greatly deteriorates the quality of the 3D portrait. The above samples fully illustrate the significance of style-directional difference loss in achieving a multi-view coherent stylization.

(a) Input (b) w/o Subspace Localization (c) w/o Directional Guidance (d) w/o Spatial Alignment (e) Full

Fig. 9: Ablation study of our three components. We showcase (b) the reconstructed 3D portrait without the content-analogous subspace localization, (c) the reconstructed 3D portrait without the style-directional guidance, and (d) the reconstructed 3D portrait without the spatial alignment loss. The reference paintings from top to bottom are *Mona Lisa*, *Red Madras Headdress*, and *Van Gogh*, respectively.

(a) Input (b) Optimization of the target view ϑ

Fig. 10: Target view optimization. (a) is the stylized input, while (b) shows the rendering of the trainable target view ϑ. The leftmost column of (b) is the initial target view, while the rightmost is the optimized view ϑ^*. The images from the left to the right show the optimization process.

(a) Input (b) Happiness (c) Contempt (d) Surprised

Fig. 11: 3D manipulation of our 3DArtmator. (a) shows the referenced portraits. (b) & (c) & (d) illustrate the 3D reconstructed portraits after editing by the expression of happiness, contempt, and surprise, respectively.

Furthermore, we explore the influence of jointly optimizing the latent code z and the user-defined target pose ϑ in Eq. (3), which can be defined as $z^*, \vartheta^* = \arg\min_{z,\vartheta} \mathcal{L}_{cr}$. Fig. 10 delineate the progession of optimizing ϑ against input artistic examples depicted on the left. The top row of Fig. 10 affirm the potential for automatic pose alignment with the given input. Nonetheless, the suboptimal poses might occur as presented in the middle and bottom rows in Fig. 10, particularly prevalent when the input involves intricate color combinations. Specifically, the bottom row exhibits a situation where the input portrait's hair is mistakenly identified as background, leading to a misjudged pose estimation. One possible explaination is that the loss $\mathcal{L}_{cr}$ excessively prioritizing the color and texture alignment over the preservation of facial orientation coherence. As a result, for the faithful reconstruction of artistic portraits, the adoption of a fixed pose ϑ is recommended.

4.6 3D Manipulation

To verify the capability of 3D artistic animation, we conduct experiments regarding the interpretability of the stylized generator. Following AniFaceGAN [55], which supports a wide range of facial attribute manipulation by alternating expression code z_{exp}, we employ three micro-expressions

(a) Input (b) Multi-view Visualization

Fig. 12: 3D reconstruction of stylized animals. (a) is the stylized input, while (b) is the multi-view results of the reconstructed 3D scenes. Three styles are watercolor, manga, and marble statue, respectively.

that involve the editing of multiple attributes. To illustrate the interpretability of our transferred style space, we manipulate the expressions by alternating the expression codes [9]. We visualize the multi-view edited portraits in Fig. 11.

The results indicate that 3DArtmator supports 3D-consistent editing after optimization. Notice that the unseen features such as the teeth in Fig. 11 are not concluded in the portrait, but still can be coherently stylized and integrated into the reconstructed portrait. This is due to the transfer of the interpretable subspace, which adapts the unseen features automatically.

Comparing the same expression manipulation on the separately optimized cases, the same expression code enables the homogeneous control, indicating that 3DArtmator preserves the interpretability of the backbone [55].

4.7 Extend to Other Domains

Ideally, 3DArtmator is not restricted to stylizing human portraits but enables generalization to other domains with the corresponding pretrained checkpoint. In Fig. 12, we extend 3DArtmator to achieve the stylized reconstruction for animal faces, with three different styles (watercolor, manga, and marble statue). We employ the weights pretrained on CATS [63]. The results in Fig. 12 demonstrate the extensibility of 3DArtmator to other categories of artforms with complex shapes, such as animal faces.

5 Conclusion

In this paper, we present 3DArtmator, a specialized 3D-aware framework designed for animating artworks. To preserve the interpretability of the pretrained generator, we utilize an anchor within the latent space, allowing for disentangment of scene reconstruction and style transfer. This approach enables 3DArtmator to achieve exceptional adaptation performance while retaining interpretability in the optimized new domain. Experimental results on diverse artforms demonstrate the effectiveness of our approach.

Limitation. While our 3DArtmator has achieved impressive results in the one-shot 3D stylization of general artistic styles, it does face limitations when it comes to stylizing more challenging styles, such as paintings with

(a) Input (b) Multi-view Visualization

Fig. 13: Failure cases. The input images are paintings of Pablo Picasso and Amedeo Modigliani.

drastic structural deviation from the canonical face. Some failure cases are shown in Fig. 13. For example, Picasso's self-portrait has oversized features and does not conform to the perspective of a real human face, thus failing to recover the obscured parts. Another example is Modigliani's portrait, which has a longer and curved nose. Instead of stretching the nose to attain a correct geometry, 3DArtmator directly maps the texture to the lip, leading to artifacts in other views. These artforms exhibit less structural similarity with the real domain, making it difficult to compensate for missing spatial information solely based on a single-view image. Besides, 3DArtmator requires instance-level optimization, which may not be as flexible as the 2D face reenactment methods.

Future works. Stylization with intricate styles remains a problem of 3DArtmator to be solved in the future. Since the features of a flat image are insufficient for fully perceiving the style in the 3D space, our future efforts will focus on introducing extra complementary information to enhance the expressiveness of styles, such as 2D landmarks supervision and stronger stylistic priors.

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM TOG*, 39(4):64–1, 2020.

[2] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pages 16123–16133, 2022.

[3] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, pages 5799–5809, 2021.

[4] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *CVPR*, pages 1897–1906, 2017.

[5] Haibo Chen, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu, et al. Artistic style transfer with internal-external learning and contrastive learning. *NeurIPS*, 34:26561–26573, 2021.

[6] Xinghao Chen, Yiman Zhang, Yunhe Wang, Han Shu, Chunjing Xu, and Chang Xu. Optical flow distillation: Towards efficient and stable video style transfer. In *ECCV*, pages 614–630. Springer, 2020.

[7] Yu-Jie Chen, Shin-I Cheng, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. Vector quantized image-to-image translation. In *ECCV*, pages 440–456, 2022.

[8] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, 2022.

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR*, pages 0–0, 2019.

[10] Zhiwen Fan, Yifan Jiang, Peihao Wang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Unified implicit neural stylization. In *ECCV*, pages 636–654. Springer, 2022.

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

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016.

[13] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *ICLR*, pages 1–25. OpenReview. net, 2022.

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

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017.

[16] Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin Gao. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *CVPR*, pages 18342–18352, 2022.

[17] Wonjong Jang, Gwangjin Ju, Yucheol Jung, Jiaolong Yang, Xin Tong, and Seungyong Lee. Stylecarigan: caricature generation via stylegan feature map modulation. *ACM TOG*, 40(4):1–16, 2021.

[18] Wonjoon Jin, Nuri Ryu, Geonung Kim, Seung-Hwan Baek, and Sunghyun Cho. Dr. 3d: Adapting 3d gans to artistic drawings. In *SIGGRAPH Asia*, pages 1–8, 2022.

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[20] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *NeurIPS*, 33:12104–12114, 2020.

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019.

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

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

[24] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *ICCV*, pages 6649–6658, 2021.

[25] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *CVPR*, pages 4990–4998, 2017.

[26] Ishit Mehta, Michaël Gharbi, Connelly Barnes, Eli Shechtman, Ravi Ramamoorthi, and Manmohan Chandraker. Modulated periodic activations for generalizable local functional representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14214–14223, 2021.

[27] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, pages 3481–3490. PMLR, 2018.

[28] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[29] Arnab Kumar Mondal, Piyush Tiwary, Parag Singla, and AP Prathosh. Few-shot cross-domain image generation via inference-time latent-code learning. In *ICLR*, 2023.

[30] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

[31] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. Snerf: stylized neural implicit representations for 3d scenes. *ACM TOG*, 41(4):1–11, 2022.

[32] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *CVPR*, pages 10743–10752, 2021.

[33] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR*, pages 13503–13513, 2022.

[34] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, pages 319–345. Springer, 2020.

[35] Justin NM Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334*, 2020.

[36] 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 *ICML*, pages 8748–8763. PMLR, 2021.

[37] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *ICCV*, pages 13759–13768, 2021.

[38] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, pages 2287–2296, 2021.

[39] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM TOG*, 42(1):1–13, 2022.

[40] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *NeurIPS*, 33:20154–20166, 2020.

[41] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. In *NeurIPS*, 2022.

[42] Min Shi, Jia-Qi Zhang, Shu-Yu Chen, Lin Gao, Y Lai, and Fang-Lue Zhang. Reference-based deep line art video colorization. *IEEE TVCG*, 20(1), 2022.

[43] Yezhi Shu, Ran Yi, Mengfei Xia, Zipeng Ye, Wang Zhao, Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. Gan-based multi-style photo cartoonization. *IEEE TVCG*, 28(10):3376–3390, 2021.

[44] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 32, 2019.

[45] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, pages 13653–13662, 2021.

[46] Guoxian Song, Hongyi Xu, Jing Liu, Tiancheng Zhi, Yichun Shi, Jianfeng Zhang, Zihang Jiang, Jiashi Feng, Shen Sang, and Linjie Luo. Agilegan3d: Few-shot 3d portrait stylization by augmented transfer learning. *arXiv preprint arXiv:2303.14297*, 2023.

[47] Haorui Song, Yong Du, Tianyi Xiang, Junyu Dong, Jing Qin, and Shengfeng He. Editing out-of-domain gan inversion via differential activations. In *ECCV*, pages 1–17, 2022.

[48] Rakshith Subramanyam, Vivek Narayanaswamy, Mark Naufel, Andreas Spanias, and Jayaraman J Thiagarajan. Improved stylegan-v2 based inversion for out-of-distribution images. In *ICML*, 2022.

[49] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM TOG*, 40(4):1–14, 2021.

[50] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Cross-domain and disentangled face manipulation with 3d guidance. *IEEE TVCG*, 29(4):2053–2066, 2022.

[51] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Nerf-art: Text-driven neural radiance fields stylization. *IEEE TVCG*, 2023.

[52] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *CVPR*, pages 11379–11388, 2022.

[53] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022.

[54] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004.

[55] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. In *NeurIPS*, 2022.

[56] Wenpeng Xiao, Cheng Xu, Jiajie Mai, Xuemiao Xu, Yue Li, Chengze Li, Xueting Liu, and Shengfeng He. Appearance-preserved portrait-to-anime translation via proxy-guided domain adaptation. *IEEE TVCG*, 2022.

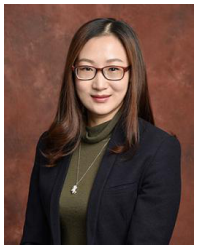
[57] Yi Xie, Hanxiao Wu, Fei Shen, Jianqing Zhu, and Huanqiang Zeng. Object re-identification using teacher-like and light students. In *BMVC*, 2021.

[58] Yi Xie, Huaidong Zhang, Xuemiao Xu, Jianqing Zhu, and Shengfeng He. Towards a smaller student: Capacity dynamic distillation for efficient image retrieval. In *CVPR*, pages 16006–

16015, 2023.
[59] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: exemplar-based high-resolution portrait style transfer. In *CVPR*, pages 7693–7702, 2022.
[60] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. *ICCV*, 2023.
[61] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *ECCV*, pages 717–733. Springer, 2022.
[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.
[63] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection-how to effectively exploit shape and texture features. In *ECCV*, pages 802–816. Springer, 2008.
[64] Yabo Zhang, Yuxiang Wei, Zhilong Ji, Jinfeng Bai, Wangmeng Zuo, et al. Towards diverse and faithful one-shot adaption of generative adversarial networks. In *NeurIPS*, 2022.
[65] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G. Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *ECCV*, 2022.
[66] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. In *ICLR*, 2022.

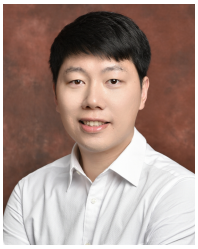
Chenxi Zheng obtained B.Sc. degree in computer science and technology from South China University of Technology in 2023. He is currently working toward the Ph.D. degree in the School of Computer Science and Engineering, South China University of Technology. His research interests include image syntesis, computer graphics, and the applications of 3D computer vision.

Bangzhen Liu obtained B.Sc. degree in computer science and technology from South China University of Technology in 2021. He is currently working toward the Ph.D. degree in the School of Computer Science and Engineering, South China University of Technology. His research interests include transfer learning, deep learning for 3D vision, and the applications of 3D computer vision.



Xuemiao Xu received her B.S. and M.S. degrees in Computer Science and Engineering from South China University of Technology in 2002 and 2005 respectively, and Ph.D. degree in Computer Science and Engineering from The Chinese University of Hong Kong in 2009. She is currently a professor in the School of Computer Science and Engineering, South China University of Technology. Her research interests include object detection, tracking, recognition, and image, video understanding and synthesis, particularly their applications in the intelligent transportation.

Huaidong Zhang is an Associate Professor in the School of Future Technology, South China University of Technology. He was a Postdoctoral Fellow at The Hong Kong Polytechnic University. He received his B.Eng. and Ph.D. degrees in Computer Science and Engineering from the South China University of Technology in 2015 and 2020 respectively. His research interests include computer vision, image processing, computer graphics and deep learning.



Shengfeng He (Senior Member, IEEE) is an associate professor in the School of Computing and Information Systems, Singapore Management University. He was on the faculty of the South China University of Technology, from 2016 to 2022. He obtained B.Sc. and M.Sc. degrees from Macau University of Science and Technology in 2009 and 2011 respectively, and a Ph.D. degree from City University of Hong Kong in 2015. His research interests include computer vision and generative models. He is a senior member of IEEE and CCF. He serves as the lead guest editor of the IJCV, the associate editor of IEEE TNNLS, IEEE TCSVT, and Neurocomputing. He also serves as the area chair/senior program committee of ICML, AAAI, IJCAI, and BMVC.