

Semantic-aware Generation of Multi-view Portrait Drawings

Biao Ma¹, Fei Gao^{2*}, Chang Jiang¹, Nannan Wang³, Gang Xu¹

¹ School of Computer Science and Technology, Hangzhou Dianzi University

² Hangzhou Institute of Technology, Xidian University

³ ISN State Key Laboratory, Xidian University

{aiartma, jc233, gxu}@hdu.edu.cn, {fgao, nnwang}@xidian.edu.cn

Abstract

Neural radiance fields (NeRF) based methods have shown amazing performance in synthesizing 3D-consistent photographic images, but fail to generate multi-view portrait drawings. The key is that the basic assumption of these methods – *a surface point is consistent when rendered from different views* – doesn't hold for drawings. In a portrait drawing, the appearance of a facial point may change when viewed from different angles. Besides, portrait drawings usually present little 3D information and suffer from insufficient training data. To combat this challenge, in this paper, we propose a *Semantic-Aware GEnerator* (SAGE) for synthesizing multi-view portrait drawings. Our motivation is that facial semantic labels are view-consistent and correlate with drawing techniques. We therefore propose to collaboratively synthesize multi-view semantic maps and the corresponding portrait drawings. To facilitate training, we design a semantic-aware domain translator, which generates portrait drawings based on features of photographic faces. In addition, use data augmentation via synthesis to mitigate collapsed results. We apply SAGE to synthesize multi-view portrait drawings in diverse artistic styles. Experimental results show that SAGE achieves significantly superior or highly competitive performance, compared to existing 3D-aware image synthesis methods. The codes are available at <https://github.com/AiArt-HDU/SAGE>.

1 Introduction

3D-aware image synthesis [Mildenhall *et al.*, 2021] aims to generate multi-view consistent images and, to a lesser extent, extract 3D shapes, without supervision on geometric or multi-view image datasets. Recently, inspired by the great success of Neural Radiation Fields (NeRF) [Mildenhall *et al.*, 2021] and Generative Adversarial Networks (GANs) [Karras *et al.*, 2019], impressive progress has been achieved in generating multi-view photos as well as detailed geometries [Deng *et al.*, 2022b; Xiang *et al.*, 2022; Chan *et al.*, 2022b]. Besides,

*Corresponding Author

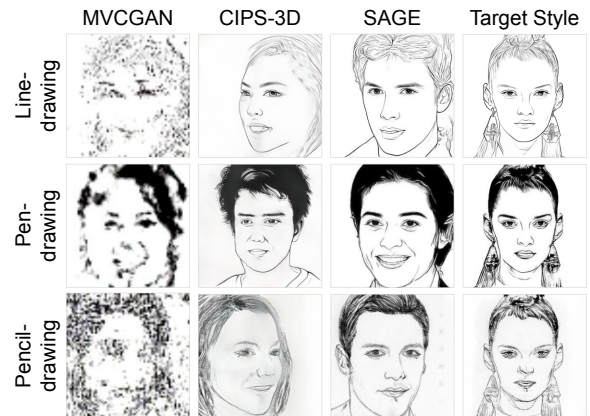


Figure 1: Portrait drawings synthesized by MVCGAN [Zhang *et al.*, 2022], CIPS-3D [Zhou *et al.*, 2021], and our method, i.e. SAGE. The final column show training examples in target styles.

several recent methods [Zhang *et al.*, 2022; Gu *et al.*, 2021; Niemeyer and Geiger, 2021; Zhou *et al.*, 2021] can also synthesize high quality artistic images, such as oil-paintings. While marveling at the impressive results of 3D-aware image synthesis methods, we wish to extend the style of synthesized images. Unfortunately, the advanced methods all fail to generate high quality multi-view portrait drawings, e.g. facial line-drawings (Fig. 1).

There are mainly three reasons for their failure. First, the assumption of NeRF-based methods – *a surface point is consistent when rendered from different views* – doesn't hold for drawings. Human artists typically use a sparse set of strokes to represent geometric boundaries, and use diverse levels of tones to present 3D structures [Sousa and Buchanan, 1999]. As illustrated in Fig. 2, the boundaries may vary when a face is viewed from different angles. In other words, the appearance of a facial point may be inconsistent between different views, in portrait drawings. Second, portrait drawings usually present sparse information with little 3D structures. Existing NeRF-based methods produce radiance fields and render images based on adjacent correlations and stereo correspondence [Zhang *et al.*, 2022; Deng *et al.*, 2022b]. As a result, it is not appropriate to directly apply previous methods for portrait drawing synthesis.

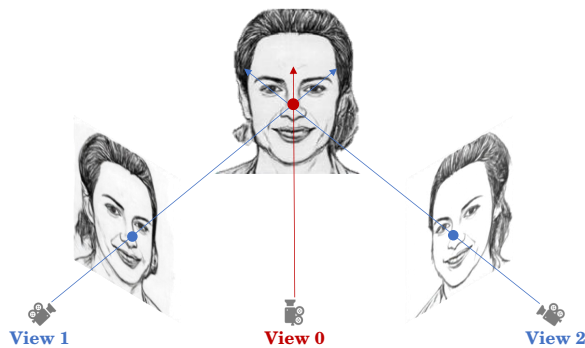


Figure 2: In portrait drawings, the appearance of a facial point may be inconsistent between different views. For example, the nose tip is represented by strokes in both View 1 and View 2, but is left blank in View 0.

Third, previous methods require a large amount of training data. Unfortunately, it is extremely time-consuming and laborious for human artists to create adequate portrait drawings.

To combat this challenge, in this paper, we propose a *Semantic-Aware GEnerator* (SAGE) for synthesizing multi-view portrait drawings. Our motivation is that facial semantic labels are view-consistent and are highly correlated with the appearance of portrait drawings. Commonly human artists draw different semantic areas by using adaptive drawing techniques [Sousa and Buchanan, 1999]. We therefore collaboratively synthesize multi-view consistent semantic maps and the corresponding portrait drawings. Besides, we use semantic maps to guide synthesis of portrait drawings, through semantic-adaptive normalization [Wang *et al.*, 2018]. As a result, the synthesized drawings are constrained to convey facial semantic structures, instead of multi-view consistency.

In addition, NeRF-based modules, including radiation fields production and volume rendering (VR) [Zhang *et al.*, 2022], are still essential for producing multi-view images. For effective learning of such modules, we propose a two-stage generation and training strategy to make NeRF-based modules suitable for synthesizing portrait drawings. First, we train the NeRF modules for generating multi-view facial photos and semantic maps in parallel. Afterwards, we use a semantic-adaptive domain translator to synthesizes portrait drawings from features of photographic faces. The warm-up in the first stage makes NeRF-based methods suitable for synthesizing portrait drawings. Consequently, the final generator becomes capable of producing high quality portrait drawings under different views. Finally, we use data augmentation via synthesis to obtain adequate training samples, and to mitigate collapsed results under large pose variants.

We apply our method, SAGE, to synthesize diverse styles of portrait drawings, including pen-drawings [Yi *et al.*, 2019], line-drawings, pencil-drawings [Fan *et al.*, 2022], and oil-paintings [Nichol, 2016]. Experimental results show that SAGE achieves significantly superior or highly competitive performance, compared to existing 3D-aware image synthesis methods. Especially, our method stably generate high quality results across a wide range of viewpoints.

2 Related Work

2.1 3D-aware Image Synthesis

3D-aware image synthesis aims to explicitly control the camera view of synthesized images. Recently, numerous NeRF-based methods have been proposed and achieved impressive progress. For example, to achieve more photorealistic results, StyleNeRF [Gu *et al.*, 2021] uses StyleGAN [Karras *et al.*, 2019] to obtain high-resolution images by up-sampling low-resolution feature maps through Convolutional Neural Networks (CNNs). StyleSDF [Or-Eli *et al.*, 2022] employs a signed distance field to obtain low-resolution spatial features and then produces high-resolution images through up-sampling. GIRAFFE [Niemeyer and Geiger, 2021] instead up-samples multi-scale feature maps based on GRAF [Schwarz *et al.*, 2020] to generate high-resolution and multi-view consistent images. Recently, MVCGAN [Zhang *et al.*, 2022] introduces a method to warp images based on camera matrices, based on pi-GAN [Chan *et al.*, 2021]. GRAM [Deng *et al.*, 2022b; Xiang *et al.*, 2022] optimizes point sampling and radiance field learning.

A similar work to ours is CIPS-3D [Zhou *et al.*, 2021], which use a NeRF-based network for view control and a 2D implicit neural representation (INR) network for generating high-resolution images. CIPS-3D is also trained in two-stages so as to synthesize artistic portraits, e.g. oil-paintings and cartoon faces. However, CIPS-3D produces geometric deformations and unpleasant artifacts, when applied to portrait drawings (Fig. 1). Differently, we propose a novel two-stage generation framework, i.e. using a domain translator to synthesize portrait drawings conditioned on photographic features. Besides, we generate multi-view images based on semantic consistency instead of 3D consistency. Recently, both FENeRF [Sun *et al.*, 2022b] and IDE-3D [Sun *et al.*, 2022a] directly use NeRF to render facial semantics, as well as multi-view images, for face editing. Differently, we decode both facial semantics and portrait drawings from photographic features, and use semantics to provide structural guidance.

2.2 2D Artistic Portrait Drawing Generation

For now, there have been a huge number of methods for generating 2D artistic portrait drawings. Most existing methods try to translate facial photos to sketches, e.g. pencil-drawings [Wang and Tang, 2009; Wang *et al.*, 2017], pen-drawings [Yi *et al.*, 2019], and line-drawings [Gao *et al.*, 2020; Yi *et al.*, 2022]. The advanced methods are mainly based on 2D conditional GANs [Isola *et al.*, 2017; Gui *et al.*, 2021] and formulate this task as image-to-image translation. Researchers have made great efforts to boost the identity-consistency and texture-realism, by designing new learning architectures [Zhang *et al.*, 2019; Zhu *et al.*, 2021; Zhang *et al.*, 2018], or using auxiliary geometric information, e.g. facial landmarks [Yi *et al.*, 2019; Gao *et al.*, 2020], semantics [Yu *et al.*, 2021; Li *et al.*, 2021], and depth [Chan *et al.*, 2022a]. There are also several unsupervised methods for synthesizing 2D portrait drawings. These methods [Noguchi and Harada, 2019; Li *et al.*, 2020; Ojha *et al.*, 2021; Kong *et al.*, 2021] aims at solving the few-shot learning problem of big generative models (e.g., StyleGAN series [Karras

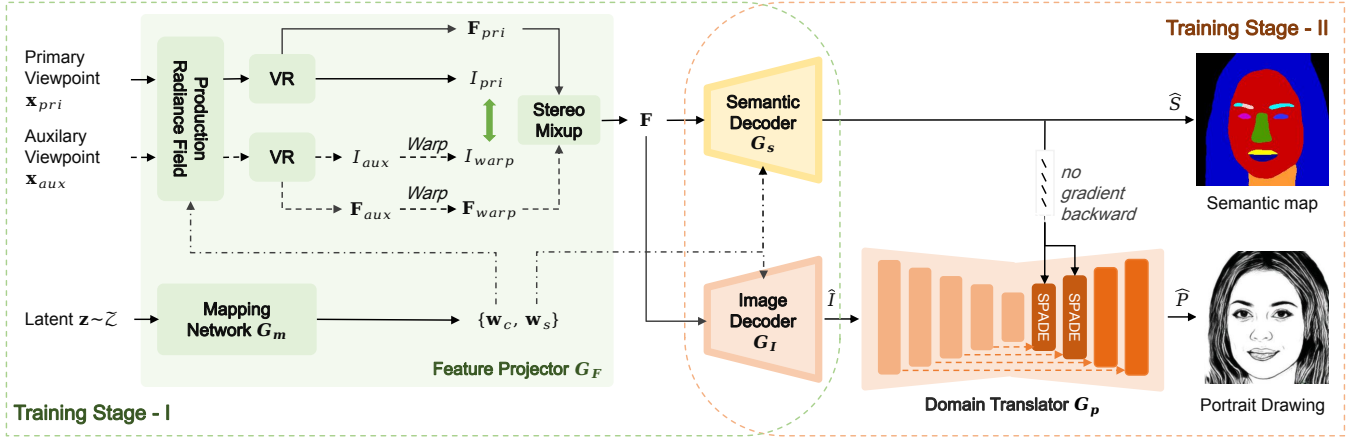


Figure 3: Pipeline of our semantic-aware portrait drawing generator, SAGE. The feature projector G_F generates feature maps \mathbf{F} and normalization parameters \mathbf{w} , which control content and viewpoints of generated faces. In Stage-I, we train the feature projector G_F , semantic decoder G_s , and image decoder G_I to enable the generator producing multi-view facial photos and semantic masks. In Stage-II, we add the portrait drawing decoder G_p , and refine all decoders for synthesizing high quality portrait drawings based on features of photographic faces.

et al., 2019]) and verify them on portrait drawings. In this paper, we use a modification of GENRE [Li *et al.*, 2021] to synthesis 2D portrait drawings for data augmentation (Section 3.4). Besides, the architecture of our domain translator is similar to GENRE (Section 3.2).

3 Method

Our whole model follows the architectures of NeRF-based GANs [Zhang *et al.*, 2022]. As shown in Fig. 3, our generator G includes four parts: a NeRF-based feature projector G_F , a semantic decoder G_s , and an image decoder G_I followed by a domain translator G_p . G_p produces portrait drawings conditioned on input latent code \mathbf{z} and viewpoint \mathbf{x} . Correspondingly, we use three discriminators to judge realism of semantic maps, facial images, and drawings, respectively.

3.1 NeRF-based Feature Projection

We first project latent codes and viewpoints to representations of multi-view faces by NeRF. Specially, we map the latent code \mathbf{z} to parameters $\mathbf{w} = \{\mathbf{w}_c, \mathbf{w}_s\}$ through a mapping network, G_m . \mathbf{w}_c controls content (e.g. identity, structure, and appearance) of the generated images. \mathbf{w}_s controls the style of synthesized portrait drawing. Given a viewpoint \mathbf{x} in space, we map it to appearance parameters through a multi-layer FiLM network [Perez *et al.*, 2018]. Afterwards, we use the *Volume Rendering* (VR) [Deng *et al.*, 2022a] module to produce a facial image and the corresponding feature map. In the implementation, \mathbf{w}_c is composed of multiple vectors representing frequencies and phases, and is fed into the FiLM layers. \mathbf{w}_s modulate deep features in semantic and image decoders, in the manner of AdaIN [Huang and Belongie, 2017].

During training, we follow the settings of MVCGAN [Zhang *et al.*, 2022]. Specially, we render two images from primary viewpoint \mathbf{x}_{pri} and auxiliary viewpoint \mathbf{x}_{aux} , respectively. The corresponding images, I_{pri} and I_{aux} , represent the same face but with different poses. Let \mathbf{F}_{pri} and \mathbf{F}_{aux} be the corresponding feature representations. Afterwards,

I_{aux}/\mathbf{F}_{aux} are geometrically aligned to I_{pri}/\mathbf{F}_{pri} through warping. The warped image I_{warp} is constrained to be the same as I_{pri} by a image-level reprojection loss (Eq.5). Besides, the primary feature \mathbf{F}_{pri} is linearly mixed with the warped feature \mathbf{F}_{warp} . The resulting mixed feature \mathbf{F} is fed into the following decoders. In the testing stage, the auxiliary position and the stereo mixup module aren't needed.

3.2 Semantic-aware Portrait Generation

To enable the network successfully produce facial structures, we propose to decode portrait drawings from features of photographic faces. Besides, we use 2D semantic information to guide the synthesis of portrait drawings.

Semantic and Image Decoders. First, we use a semantic G_s and an image decoder G_I to collaboratively synthesize semantic maps and facial photos. Both decoders follow the same architectures, but with different numbers of output channels. As shown in Fig. 4, each decoder contains three upsampling convolutional layers, and progressively generates high-resolution outputs. As previously mentioned, features over each layer are channel-wisely modulated by \mathbf{w}_s in the manner of AdaIN. Besides, we produce multi-scale outputs and integrate them together for producing a final output. The image decoder, G_I , generates a RGB facial image, i.e.

$$\hat{\mathbf{I}}^{3 \times 256 \times 256} = G_I(\mathbf{F}, \mathbf{w}_s). \quad (1)$$

The semantic decoder, G_s , produces the corresponding 19-channel semantic masks [Lee *et al.*, 2020], i.e.

$$\hat{\mathbf{S}}^{19 \times 256 \times 256} = G_s(\mathbf{F}, \mathbf{w}_s). \quad (2)$$

Semantic-adaptive Domain Translator. In addition, we use a semantic-adaptive domain translator, G_p , to generate portrait drawings based on features of photographic faces. We design G_p following the structure of U-Net [Isola *et al.*, 2017], and use facial semantics $\hat{\mathbf{S}}$ to guide synthesis in the manner of SPADE [Park *et al.*, 2019; Li *et al.*, 2021]. Since our model is trained without supervision, the synthesized

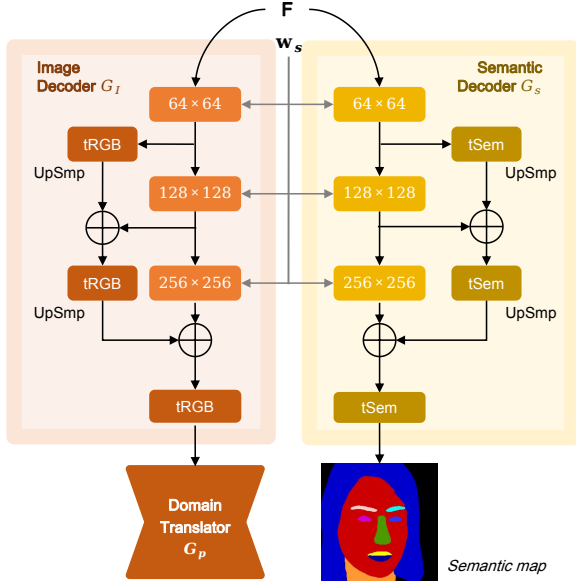


Figure 4: Semantic decoder (right) and portrait decoder (left and bottom). The tRGB and tSem module map deep features to a RGB image and semantic maps by 1×1 convolution, respectively. UpSmp denotes upsampled interpolation.

semantic maps might be incorrect. We therefore only use SPADE over small-scale layers in G_p to control the major structure of portrait drawings. The process of portrait drawing synthesis is formulated as:

$$\hat{P}^{3 \times 256 \times 256} = G_p(\hat{I}, \hat{S}). \quad (3)$$

The semantic guidance will enable our model producing natural and distinct facial structures, as well as semantic-related details (Section 4.4). In addition, with the semantic modulation module, our method allows minor editing on portrait drawings (Section 4.5).

3.3 Two-stage Training

In preliminary experiments, we train the whole model end-to-end on portrait drawings. The model produces messy ink patterns, and fails to generate acceptable facial structures (Section 4.4). As previously analyzed, the possible reason is that portrait drawings present sparse appearances with little 3D information. To combat this challenge, we propose to train the model in two stages. First, we train the feature projector, semantic decoder, and image decoder, to enable them generating high quality facial structures. Afterwards, we add the domain translator G_p , and refine all the decoding modules for synthesizing multi-view portrait drawings. This training strategy conveys well with the architecture of our generator.

Training Stage-I. In the first training stage, we use facial photos and their corresponding semantic maps [Lee *et al.*, 2020] to train our model without G_p . We use discriminators similar to pi-GAN [Chan *et al.*, 2021], and use a semantic discriminator D_s and an image discriminator D_I during training.

The loss function of the discriminators are defined as:

$$\begin{aligned} \mathcal{L}_{D_s} &= \mathbb{E}_{S \sim \mathcal{S}} [f(D_s(S)) + \lambda_1 |\nabla D_s(S)|^2] \\ &\quad + \mathbb{E}_{z \sim \mathcal{Z}, x \sim \mathcal{X}} [f(-D_s(\hat{S}))], \\ \mathcal{L}_{D_I} &= \mathbb{E}_{I \sim \mathcal{I}} [f(D_I(I)) + \lambda_1 |\nabla D_I(I)|^2] \\ &\quad + \mathbb{E}_{z \sim \mathcal{Z}, x \sim \mathcal{X}} [f(-D_I(\hat{I}))], \end{aligned} \quad (4)$$

where $f(u) = -\log(1 + \exp(-u))$; I and S denote facial photos and semantic maps in the training set $\{\mathcal{I}, \mathcal{S}\}$. The loss function of the generator (without G_p) is:

$$\begin{aligned} \mathcal{L}_G^{(1)} &= \mathbb{E}_{z \sim \mathcal{Z}, x \sim \mathcal{X}} [f(-D_s(\hat{S})) + f(-D_I(\hat{I}))] + \lambda_2 \mathcal{L}_{rec}, \\ \mathcal{L}_{rec} &= \lambda_3 |I_{pri} - I_{aux}| + (1 - \lambda_3) \text{SSIM}(I_{pri}, I_{aux}), \end{aligned} \quad (5)$$

where SSIM denotes the structural similarity [Wang *et al.*, 2004] between I_1 and I_2 ; $\lambda_{1,2,3}$ are weighting factors.

Training Stage-II. In the second stage, we load the pre-trained model and add the portrait drawing decoder G_p to it. Afterwards, we fix parameters of G_F and refine all the other parts, by using portrait drawings and their semantic maps $\{P \sim \mathcal{P}, S \sim \mathcal{S}\}$. Here, we use an portrait drawing discriminator D_p and the previous semantic discriminator D_s during training. The loss functions of D_p and G are defined as:

$$\begin{aligned} \mathcal{L}_{D_p} &= \mathbb{E}_{P \sim \mathcal{P}} [f(D_p(P)) + \lambda_1 |\nabla D_p(P)|^2] \\ &\quad + \mathbb{E}_{z \sim \mathcal{Z}, x \sim \mathcal{X}} [f(-D_p(\hat{P}))], \\ \mathcal{L}_G^{(2)} &= \mathbb{E}_{z \sim \mathcal{Z}, x \sim \mathcal{X}} [f(-D_s(\hat{S})) + f(-D_p(\hat{P}))]. \end{aligned} \quad (6)$$

We remove the image discriminator D_I and \mathcal{L}_{rec} in this stage, so as to eliminate their interference on synthesizing realistic portrait drawings.

3.4 Data Augmentation via Synthesis

To effectively learn a 3D-aware image synthesis model, it's critical to collect a huge number of training images of the same style. This requirement is almost impossible in our task. It's time-consuming and laborious for human artists draw adequate portraits. In existing portrait drawing datasets [Fan *et al.*, 2022; Yi *et al.*, 2019], there are merely hundreds of samples for a given style. These observations imply the significance of data augmentation for training. In the implementation, we train a modification of GENRE [Li *et al.*, 2021] to synthesize adequate 2D portrait drawings. Afterwards, we use the synthesized data to train our model in Stage-II.

4 Experiments and Analysis

4.1 Settings

To verify and analyze our method, we conduct thorough experiments by using the following datasets: **Pen-drawings.** We conduct experiments on the APDrawing [Yi *et al.*, 2019] dataset, which contains 70 pen-drawings of faces. **Line-drawings.** We drew about 800 facial line-drawings ourselves. **Pencil-drawings.** We apply our method to the three styles of pencil-drawings on the FS2K [Fan *et al.*, 2022] dataset. Each style contains about 350 training images. **Oil paintings.** We randomly select 3,133 oil-paintings of humans from WikiArt

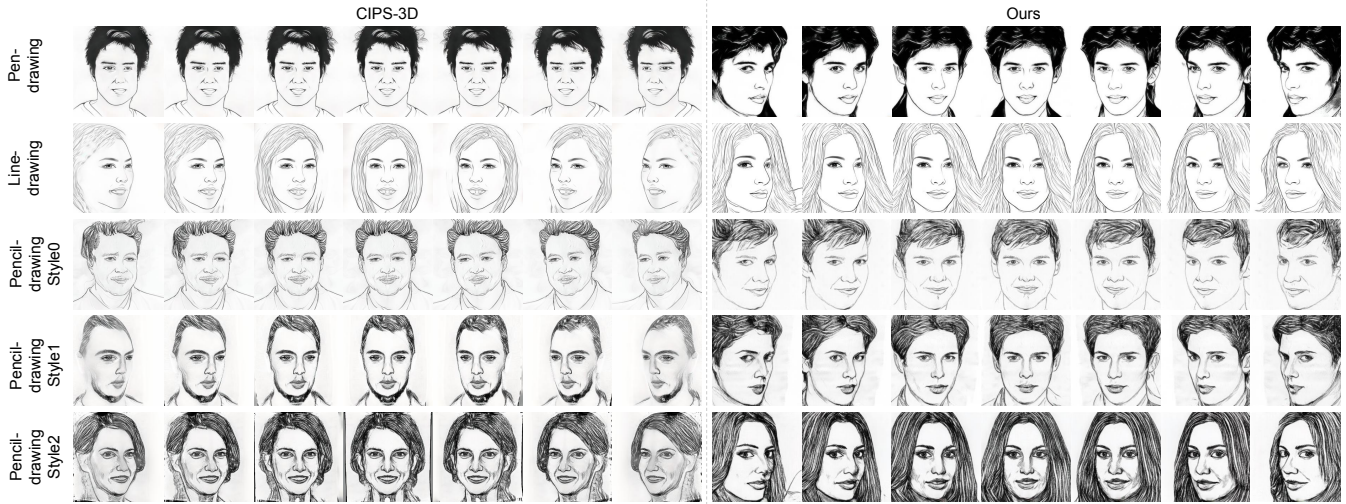


Figure 5: Qualitative comparison on Pen-drawings, Line-drawings, and three styles of Pencil-drawings [Fan *et al.*, 2022], at 256^2 resolution. The left part show results of CIPS-3D [Zhou *et al.*, 2021], the right part show ours.

[Nichol, 2016] to evaluate the generalization capability of our model. **Facial Photos.** We use the CelebAMask-HQ [Lee *et al.*, 2020] dataset during training stage-I and for data augmentation. It contains 30,000 facial photos and the corresponding semantic maps. For Pen-drawings, Line-drawings, and three styles of Pencil-drawings, we train a 2D GAN by using the corresponding real portrait drawings. Afterwards, we apply the learned models to the facial photos in CelebAMask-HQ dataset. Finally, the synthesized portrait drawings are used to train 3D-aware image synthesis models.

4.2 Training Details

In the pre-training phase of the model, we set $\lambda_1 = 0.1$, $\lambda_2 = 1$, and $\lambda_3 = 0.25$. In the second phase of training, the parameters remain unchanged. We use the Adam optimizer and set $\beta_1 = 0$ and $\beta_2 = 0.9$. At 64^2 resolution, the learning rate of generator is 6×10^{-5} , the learning rate of discriminators 2×10^{-4} , and batch size 36. At 128^2 resolution, the learning rate of generator is 5×10^{-5} and batch size 24. At 256^2 resolution, the learning rate of generator is 3×10^{-5} , the learning rate of discriminators 1×10^{-4} , and batch size 24. We use a single GeForce RTX3090 to train our model.

4.3 Comparison with SOTAs

We compare SAGE with two SOTA methods, i.e. CIPS-3D [Zhou *et al.*, 2021] and MVCGAN [Zhang *et al.*, 2022]. We use their official implementations provided by the authors, and conduct experiments under the same settings as ours.

Qualitative comparison on portrait-drawings. As previously presented in Fig. 1, MVCGAN fails to generate acceptable portrait drawings. Fig. 5 shows the portrait drawings generated by CIPS-3D and SAGE. Here, we show seven views of each sample. Obviously, our results show more natural structures than CIPS-3D. The synthesized drawings of CIPS-3D, especially the pen-drawings and pencil-drawings, present geometric deformations. Besides, CIPS-3D produces unappealing artifacts under large pose variants. In contrast,

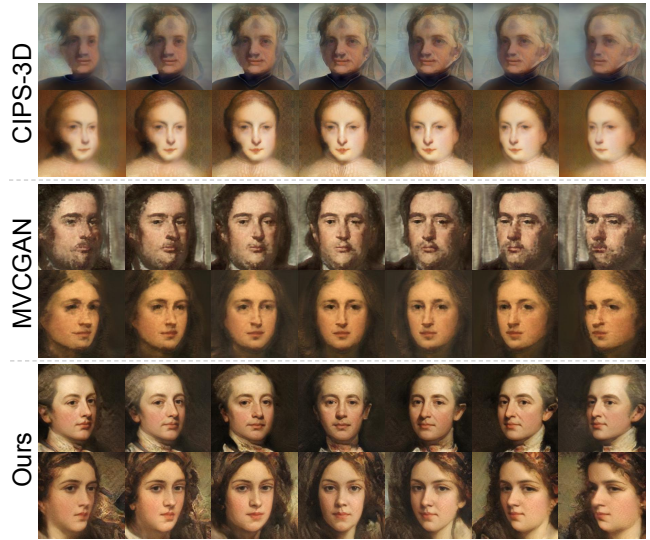


Figure 6: Qualitative comparison in synthesizing multi-view oil-paintings at 256^2 resolution.

our method, SAGE, generates high quality portrait drawings across all the poses and all the styles. Specially, our synthesized portrait drawings present distinct facial structures with realistic strokes and textures.

Quantitative comparison on portrait-drawings. We further evaluate the quality of synthesized images by using the Fréchet Inception Distances (FID) [Heusel *et al.*, 2017] and sliced Wasserstein distance (SWD) [Karras *et al.*, 2017]. For each style of portrait drawing, we use $4K$ generated images and $4K$ real images for computing FID and SWD. As shown in Table1, our method achieves significantly lower values of both FID and SWD than CIPS-3D and MVCGAN; except that the SIFID value is slightly higher than CIPS-3D on

	Pen-drawing		Line-drawing		Pencil-drawing Style0		Pencil-drawing Style1		Pencil-drawing Style2		Oil-paintings	
	SIFID	SWD	SIFID	SWD	SIFID	SWD	SIFID	SWD	SIFID	SWD	SIFID	SWD
CIPS-3D [Zhou <i>et al.</i> , 2021]	3.51	86.9	3.31	91.3	5.49	40.7	5.12	25.1	5.33	24.9	4.86	14.0
MVCGAN [Zhang <i>et al.</i> , 2022]	4.19	59.8	3.69	99.1	5.76	48.6	5.15	24.7	5.29	25.5	4.56	36.3
SAGE (Ours)	3.07	38.5	3.04	89.8	5.16	34.0	5.13	19.7	5.11	19.9	4.71	26.5

Table 1: Quantitative comparison between SAGE and advanced methods, including CIPS-3D [Zhou *et al.*, 2021] and MVCGAN [Zhang *et al.*, 2022]. Smaller values of both SIFID and SWD indicate better quality of synthesised images. The best criteria are highlighted in **bold**.

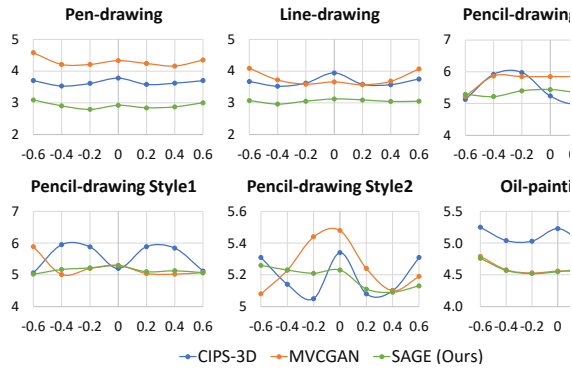


Figure 7: Curves of SIFID values across different views. T_i denotes the pose variants sequentially corresponding to F_i Fig. 6. The y -axis denotes values of SIFID.

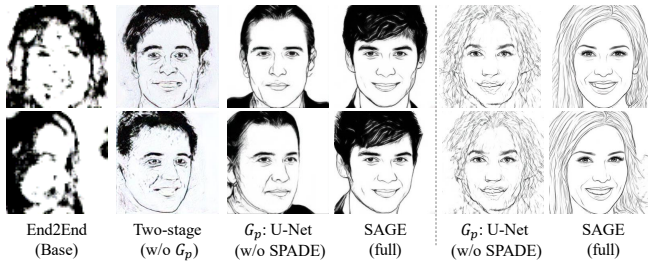


Figure 8: Results of ablation study.

Pencil-drawing Style 1. Such comparison results are consistent with Fig. 5, and further demonstrate the superiority of our method in generating multi-view portrait drawings.

Comparison on oil-paintings. Since both MVCGAN and CIPS-3D perform well on synthesizing oil-paintings, we additionally compare with them on the WikiArt dataset. Fig. 6 shows the results of multi-view oil-paintings synthesis. Obviously, CIPS-3D produces geometric deformations and unappealing artifacts. MVCGAN produces reasonable paintings but with blurring details. In contrast, our results present the best quality, in terms of either facial structures or painting textures. Table 1 shows the SIFID and SWD values computed from 1500 generated paintings and 1500 real ones. Our method achieves the best SIFID value, but shows inferiority to CIPS-3D in terms of SWD. Based on all the comparison results, we can safely conclude that our method is superior and at least highly competitive to SOTAs in synthesizing multi-view oil-paintings.

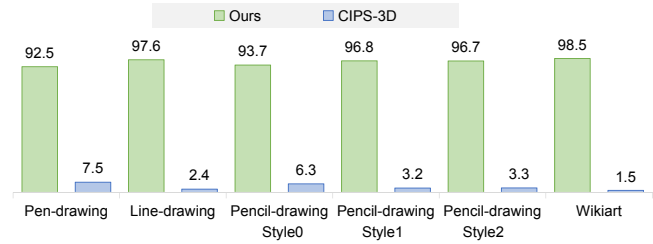


Figure 9: The average preference percent of CIPS-3D and our method w.r.t. each style of artistic portrait drawings.

Stability across pose variants. We finally analyze the stability of the three models in generating multi-view images. Specially, we compute the SIFID values corresponding to different viewpoints/poses. Fig. 7 shows that the SIFID values of either CIPS-3D or MVCGAN present dramatically fluctuate in most cases. Such fluctuations indicate the quality of generated portrait drawings changes dramatically with pose variants. In contrast, our method achieves the best and most stable performance in general, across all the six styles. In other words, our method consistently produces high quality multi-view portrait drawings. This comparison result is consistent with Fig. 5 and Fig. 6.

User study. We conducted a series of user study. Specially, we adopt the stimulus-comparison method following ITU-R BT.500-12. We randomly select 600 generated comparison pairs in total, i.e. 100 pairs for each style. Each pair is evaluated by 37 participants. Finally, we collect 22,200 preference labels. Fig.9 shows the average preference percent of CIPS-3D and SAGE, w.r.t. each style. Obviously, over 92% of participants think the portraits generated by SAGE are better than CIPS-3D, across all the styles. Such results demonstrate that our method is significantly better than CIPS-3D in synthesizing multi-view artistic portrait drawings.

4.4 Ablation Study

We conduct a series of ablation experiments to validate the significance of our proposed techniques, including two-stage training, the domain translator G_p , and the use of semantic masks, i.e. SPADE. We build several model variants by progressively using these techniques on our base model. The experiments are mainly conducted on the APDrawing dataset. The corresponding results are shown in Fig. 8 and Table 2.

Two-stage training strategy. We compare end-to-end training (End2End) with two-stage training on the base model. As shown in Fig. 8, end-to-end training leads to messy facial structures. In contrast, our two-stage training

Training		G_p			
End2End	Two-Stage	U-Net	SPADE	SIFID	SWD
✓				7.36	92.78
	✓			4.29	65.83
	✓	✓		3.18	43.27
	✓	✓	✓	3.07	38.45

Table 2: Quantitative results of ablation experiments about the training strategy and the domain translator G_p , on pen-drawing synthesis.

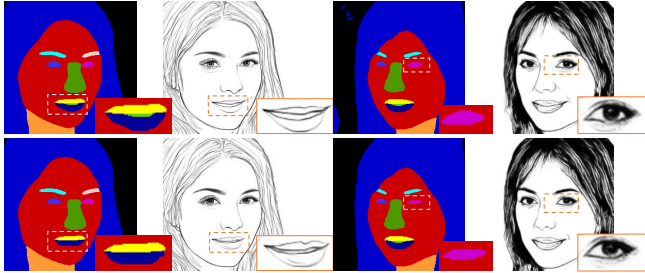


Figure 10: Semantic editing on synthesized portrait drawings.

enables the base model producing natural facial structures. Besides, two-stage training dramatically decreases the SIFID and SWD values. Such significant performance improvement demonstrate our motivation of using two-stage training.

Domain translator G_p . We further evaluate the role of domain translator. We here build a model variant by removing the SPADE modules from G_p . the architecture of pristine U-Net. As shown in Fig. 8, using a domain translator, even U-Net, dramatically boosts the quality of generated pen-drawings. Correspondingly, Table 2 shows that using U-Net alone significantly decreases both SIFID and SWD values. These results verify our motivation of using a domain translator to decode portrait drawings from features of facial photos. In this way, our generator can produce distinct facial structures and realism artistic patterns.

Semantic guidance. Finally, we analyze the role of semantics in G_p . As shown in Fig.8, SPADE improves clarity and continuity of facial semantic boundaries. Correspondingly, SPADE further decreases SIFID and SWD; and our full model achieves the best performance among all the model variants. To further verify the significance of facial semantics, we conduct experiments on line-drawing synthesis. The last two columns of Fig.8 show the corresponding results. Obviously, without the guidance of facial semantics, the model generate chaotic lines. In contrast, our full model generate distinctly better line-drawings.

These results demonstrate that semantic guidance are crucial for generating multi-style portrait drawings. From one hand, semantic masks enable the domain translator producing distinct facial structures. For the other hand, facial semantics are highly correlated with the drawing techniques, human artists used during the creation process.

4.5 Applications

Semantic editing. Recall that we use semantic maps in the domain translator G_p to control facial structures of portrait

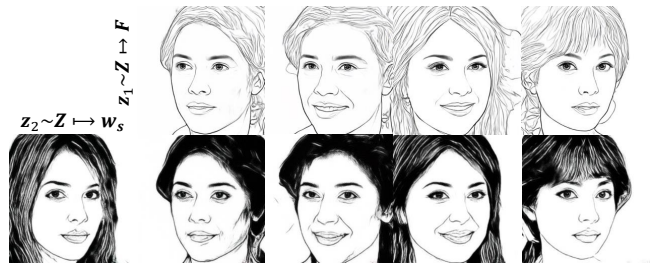


Figure 11: Style transfer from line-drawings to pen-drawings.



Figure 12: Result of identity interpolation.

drawings. As a result, our model allows for editing of portrait drawings, to a certain extent. In other words, if we modify the input semantics of G_p slightly, the synthesized portrait drawing will change accordingly. As shown in Fig. 10, when we change the semantic masks of teeth or eyes, the corresponding areas in synthesized images change accordingly.

Style transfer. In our generator, the content and style information are disentangled and stored in F and w_s , respectively. It is thus possible for us to change a portrait drawing to another style without changing the facial content. To this end, we put a latent code z_1 into the learned *line-drawing* model to get the content feature F ; and put another latent code z_2 into the learned *pen-drawing* model to get the style vector w_s . Afterwards, we use F and w_s for decoding a *pen-drawing*. As shown in Fig. 11, the line-drawings are transferred to pen-drawings, while preserving facial semantic content.

Identity interpolation. We also perform identity interpolation experiments on SAGE. Given two synthesized images, we perform linear interpolation in the content space $w_c \sim \mathcal{W}_c$ and viewpoint space $x \sim \mathcal{X}$. Fig. 12 shows interpolation results on Pen-drawings. The smooth transition in pose and appearance implies that SAGE allows precise control on both facial identity and pose.

5 Conclusion

We propose a novel method, SAGE, for generating multi-view portrait drawings. SAGE is designed to collaboratively synthesize facial photos, semantic masks, and portrait drawings. Extensive experiments and a series of ablation study are conducted on six styles of artistic portraits. SAGE stably shows impressive performance in generating high quality portrait drawings, and outperforms previous 3D-aware image synthesis methods. In the future, we are interested in synthesizing multi-view portrait drawings conditioned on a single photo. One possible way is to incorporate GAN-inversion and few-shot learning techniques.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103202; in part by the Fundamental Research Funds for the Provincial Universities of Zhejiang GK239909299001-411; in part by the National Natural Science Foundation of China under Grants 61971172, 61971339, 62176230, U22A2033, U22A2096; in part by the Zhejiang Provincial Science and Technology Program in China under Grant 2021C01108; in part by the NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization under Grant U1909210; in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15; and in part by Open Research Projects of Zhejiang Lab under Grant 2021KG0AB01.

References

- [Chan *et al.*, 2021] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021.
- [Chan *et al.*, 2022a] Caroline Chan, Frédo Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7915–7925, 2022.
- [Chan *et al.*, 2022b] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.
- [Deng *et al.*, 2022a] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022.
- [Deng *et al.*, 2022b] Yu Deng, Jiaolong Yang, Jianfeng Xi-ang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10673–10683, 2022.
- [Fan *et al.*, 2022] Deng-Ping Fan, Ziling Huang, Peng Zheng, Hong Liu, Xuebin Qin, and Luc Van Gool. Facial-sketch synthesis: A new challenge. *Machine Intelligence Research*, 19(4):257–287, 2022.
- [Gao *et al.*, 2020] Fei Gao, Jingjie Zhu, Zeyuan Yu, Peng Li, and Tao Wang. Making robots draw a vivid portrait in two minutes. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9585–9591. IEEE, 2020.
- [Gu *et al.*, 2021] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021.
- [Gui *et al.*, 2021] J Gui, Z Sun, Y. Wen, D. Tao, and J. Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–28, 2021.
- [Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [Huang and Belongie, 2017] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.
- [Isola *et al.*, 2017] 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.
- [Karras *et al.*, 2017] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- [Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405. IEEE, 2019.
- [Kong *et al.*, 2021] Chaerin Kong, Jeessoo Kim, Donghoon Han, and Nojun Kwak. Smoothing the generative latent space with mixup-based distance learning. *arXiv preprint arXiv:2111.11672*, 2021.
- [Lee *et al.*, 2020] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [Li *et al.*, 2020] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020.
- [Li *et al.*, 2021] Xiang Li, Fei Gao, and Fei Huang. High-quality face sketch synthesis via geometric normalization and regularization. In *IEEE International Conference on Multimedia and Expo (ICME) 2021*, July 5-9 2021.
- [Mildenhall *et al.*, 2021] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [Nichol, 2016] K Nichol. Painter by numbers, wikiart (2016). URL <https://www.kaggle.com/c/painter-by-numbers>, 2016.

- [Niemeyer and Geiger, 2021] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.
- [Noguchi and Harada, 2019] Atsuhiko Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2750–2758, 2019.
- [Ojha *et al.*, 2021] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021.
- [Or-El *et al.*, 2022] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022.
- [Park *et al.*, 2019] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.
- [Perez *et al.*, 2018] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [Schwarz *et al.*, 2020] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020.
- [Sousa and Buchanan, 1999] Mario Costa Sousa and John W Buchanan. Computer-generated graphite pencil rendering of 3d polygonal models. In *Computer Graphics Forum*, volume 18, pages 195–208. Wiley Online Library, 1999.
- [Sun *et al.*, 2022a] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *arXiv e-prints*, pages arXiv–2205, 2022.
- [Sun *et al.*, 2022b] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2022.
- [Wang and Tang, 2009] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE TPAMI*, 31(11):1955–1967, 2009.
- [Wang *et al.*, 2004] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [Wang *et al.*, 2017] N. Wang, S. Zhang, C. Peng, J. Li, and X. Gao. *Face Sketch Recognition via Data-Driven Synthesis*. 2017.
- [Wang *et al.*, 2018] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.
- [Xiang *et al.*, 2022] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. *arXiv preprint arXiv:2206.07255*, 2022.
- [Yi *et al.*, 2019] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2019.
- [Yi *et al.*, 2022] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Quality metric guided portrait line drawing generation from unpaired training data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:905–918, 2022.
- [Yu *et al.*, 2021] J. Yu, S. Shi, F. Gao, D. Tao, and Q. Huang. Towards realistic face photo-sketch synthesis via composition-aided GANs. *IEEE TCYB*, 51(9):4350–4362, 2021.
- [Zhang *et al.*, 2018] S. Zhang, R. Ji, J. Hu, Y. Gao, and Lin C.-W. Robust face sketch synthesis via generative adversarial fusion of priors and parametric sigmoid. In *IJCAI*, pages 1163–1169, 2018.
- [Zhang *et al.*, 2019] M. Zhang, R. Wang, X. Gao, J. Li, and D. Tao. Dual-transfer face sketch-photo synthesis. *TIP*, 28(2):642–657, Feb 2019.
- [Zhang *et al.*, 2022] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Pan Pan, and Yi Yang. Multi-view consistent generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18450–18459, 2022.
- [Zhou *et al.*, 2021] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv e-prints*, pages arXiv–2110, 2021.
- [Zhu *et al.*, 2021] Mingrui Zhu, Changcheng Liang, Nannan Wang, Xiaoyu Wang, Zhifeng Li, and Xinbo Gao. A sketch-transformer network for face photo-sketch synthesis. In *International Joint Conference on Artificial Intelligence*, 2021.