Learning Detailed Radiance Manifolds for High-Fidelity and 3D-Consistent Portrait Synthesis from Monocular Image

Yu Deng   Baoyuan Wang   Heung-Yeung Shum

Xiaobing.AI

Figure 1. Novel view synthesis results by our method. It can generate novel views of a portrait image with high-fidelity and strong 3D-consistency via single forward pass (*e.g.*, see the bangs and wrinkles). **Best viewed with zoom-in and see the video for animations.**

Abstract

A key challenge for novel view synthesis of monocular portrait images is 3D consistency under continuous pose variations. Most existing methods rely on 2D generative models which often leads to obvious 3D inconsistency artifacts. We present a 3D-consistent novel view synthesis approach for monocular portrait images based on a recent proposed 3D-aware GAN, namely Generative Radiance Manifolds (GRAM) [13], which has shown strong 3D consistency at multiview image generation of virtual subjects via the radiance manifolds representation. However, simply learning an encoder to map a real image into the latent space of GRAM can only reconstruct coarse radiance manifolds without faithful fine details, while improving the reconstruction fidelity via instance-specific optimization is time-consuming. We introduce a novel detail manifolds reconstructor to learn 3D-consistent fine details on the radiance manifolds from monocular images, and combine them with the coarse radiance manifolds for high-fidelity reconstruction. The 3D priors derived from the coarse radiance manifolds are used to regulate the learned details to ensure reasonable synthesized results

at novel views. Trained on in-the-wild 2D images, our method achieves high-fidelity and 3D-consistent portrait synthesis largely outperforming the prior art. Project page:
https://yudeng.github.io/GRAMInverter/

1. Introduction

Synthesizing photorealistic portrait images of a person from an arbitrary viewpoint is an important task that can benefit diverse downstream applications such as virtual avatar creation and immersive online communication. Thanks to the thriving of 2D Generative Adversarial Networks (GANs) [18, 25, 26], people can now generate high-quality portraits at desired views given only monocular images as input, via a simple *invert-then-edit* strategy by conducting GAN inversion [1, 38, 52] and latent space editing [12, 20, 41, 50]. However, existing 2D GAN-based methods still have deficiencies when applied to applications that require more strict 3D consistency (*e.g.* VR&AR). Due to the non-physical rendering process of the 2D CNN-based generators, their synthesized images under pose changes usually bear certain kinds of multiview inconsistency, such

as geometry distortions [3,5] and texture sticking or flickering [24,55]. These artifacts may not be significant enough when inspecting each static image but can be easily captured by human eyes under continuous image variations.

Recently, there are an emerging group of 3D-aware GANs [9,10,13,19,39,40] targeting at image generation with 3D pose disentanglement. By incorporating Neural Radiance Field (NeRF) [29] and its variants into the adversarial learning process of GANs, they can produce realistic images with strong 3D-consistency across different views, given only a set of monocular images as training data. As a result, 3D-aware GANs have shown greater potential than 2D GANs for pose manipulations of portraits. However, even though 3D-aware GANs are capable of generating 3D-consistent portraits of virtual subjects, leveraging them for real image pose editing is still a challenging task. To obtain faithful reconstructions of real images, most existing methods [9,10,13,27,46,47,67] turn to a time-consuming and instance-specific optimization to invert the given images into the latent space of a pre-trained 3D-aware GAN, which is hard to scale-up. And simply enforcing an encoder-based 3D-aware GAN inversion [7,46] often fails to preserve fine details in the original image.

In this paper, we propose a novel approach *GRAMInverter*, for high-fidelity and 3D-consistent novel view synthesis of monocular portraits via single forward pass. Our method is built upon the recent GRAM [13] that can synthesize high-quality virtual images with strong 3D consistency via the radiance manifolds representation [13]. Nevertheless, GRAM suffers from the same lack-of-fidelity issue when combined with a general encoder-based GAN inversion approach [52]. The main reason is that the obtained semantically-meaningful low-dimensional latent code cannot well record detail information of the input, as also indicated by some recent 2D GAN inversion methods [52,55].

To tackle this problem, our motivation is to further learn 3D-space high-frequency details and combine them with the coarse radiance manifolds obtained from the general encoder-based inversion of GRAM, to achieve faithful reconstruction and 3D-consistent view synthesis. A straightforward way to achieve this is to extract a high-resolution 3D voxel from the input image and combine it with the coarse radiance manifolds. However, this is prohibited by modern GPUs due to the high memory cost of the 3D voxel. To tackle this problem, we turn to learn a high resolution detail manifolds, taking the advantage of the radiance manifolds representation of GRAM, instead of learning the memory-consuming 3D voxel. We introduce a novel detail manifolds reconstructor to extract detail manifolds from the input images. It leverages manifold super-resolution [60] to predict high-resolution detail manifolds from a low resolution feature voxel. This can be effectively achieved by a set of memory-efficient 2D convolution blocks. The obtained

high resolution detail manifolds can still maintain strict 3D consistency due to lying in the 3D space. We also propose dedicated losses to regulate the detail manifolds via 3D priors derived from the coarse radiance manifolds, to ensure reasonable novel view results.

Another contribution of our method is an improvement upon the memory and time-consuming GRAM, without which it is difficult to be integrated into our GAN inversion framework. We replace the original MLP-based radiance generator [13] in GRAM with a StyleGAN2 [26]-based triplane generator proposed by [9]. The efficient GRAM requires only $1/4$ memory cost with $7\times$ speed up, without sacrificing the image generation quality and 3D consistency.

We train our method on FFHQ dataset [25] and conduct multiple experiments to demonstrate its advantages on pose control of portrait images. Once trained, GRAMInverter takes a monocular image as input and predicts its radiance manifolds representation for novel view synthesis at 3 FPS on a single GPU. The generated novel views well preserve fine details in the original image with strong 3D consistency, outperforming prior art by a large margin. We believe our method takes a solid step towards efficient 3D-aware content creation for real applications.

2. Related Work

3D-aware generative model. Learned with monocular 2D images, 3D-aware GANs [9,10,13,17,19,30,31,34,39,40,45,60,62] achieve an explicit disentanglement of camera pose by introducing underlying 3D representations. Earlier works [30,42,48] utilize voxel or mesh as the intermediate representation. Later works [9,10,13,34,39,62,69] leverage NeRF [29] and its variants [13,33,54,64] to achieve more strict 3D consistency. Among them, methods that directly render their 3D representations for image synthesis achieve the best multiview consistency [10,13,17,40,45]. We propose a novel approach for high-quality pose editing of given portraits based on GRAM [13], which is a recent 3D-aware GAN with state-of-the-art multiview consistency.

GAN inversion. GAN inversion aims to map a given real image into the latent space of a pre-trained generator for image reconstruction and manipulation. Numerous methods [1,4,6,26,37,38,52,55,63,72,73] try to find a latent code which can faithfully reconstruct the given image meanwhile falls inside a semantically meaningful latent space that supports reasonable editing. They either adopt optimization-based approach [1,2,26,73], introduce an extra image encoder [4,37,52], or utilize a hybrid version of the former two [6,72]. Nevertheless, recent studies [38,52,55,74] reveal that it is difficult to achieve high-fidelity reconstruction and artifacts-free editing at once given only low-bitrate latent code as representation. As a result, several methods propose to further fine-tune the pre-trained generator [38]
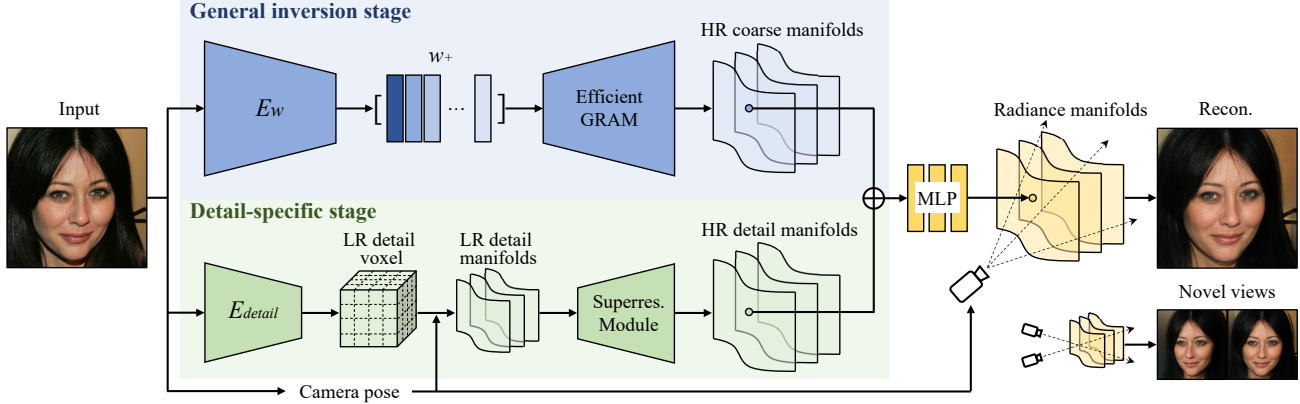
Figure 2. Overview of the GRAMInverter. An input portrait image goes through two stages to obtain the final radiance manifolds for novel view synthesis. The first general inversion stage maps the input image to the latent space of a pre-trained efficient GRAM to obtain coarse radiance manifolds. The second detail-specific stage then extracts detail feature manifolds from the input image and combines them with the coarse results for high-fidelity image synthesis. See the text for more details.

or allow more detailed features from the input images to leak into the generator during inversion [55, 63].

While the above methods target at the inversion of 2D GANs, inverting a given image with 3D-aware GANs shares a similar spirit. An advantage of 3D-aware GAN inversion compare to its 2D counterpart is a natural disentanglement of the 3D pose. Once inverted, novel view synthesis can be easily achieved without further latent space exploration [20, 41]. The majority of existing 3D-aware GAN inversion methods [9,10,13,27,46,47,51,57] leverage optimization-based or a hybrid approach for faithful reconstruction, which are time-consuming and hard to scale-up. A recent method [7] explores GAN inversion with a single forward of an encoder, yet it struggles to preserve fine image details. Our proposed method is also an encoder-based inversion approach which yields high-quality reconstruction and novel view synthesis thanks to our novel design.

Pose editing of monocular portraits. Editing the camera pose of a monocular portrait for novel view synthesis is a longstanding task and has witnessed the emergence of diverse methods. Some of them [8, 35, 59, 61, 75, 76] achieve pose editing by first conducting 3D reconstruction and then rendering the obtained mesh at novel views. Due to the imperfect reconstruction results, they often have difficulties handling non-face regions and unseen parts at the input view. Others [15, 16, 32, 36, 43, 44, 56, 58, 66, 70] generate novel views in a face-reenactment paradigm, where warping flows are often learned from video data to transform a source image to a target viewpoint. These methods may encounter geometry distortions at novel views due to the lack of an explicit 3D constraint. More recently, plenty of works [3,7,12,22,28,41,46,47,50] have studied pose editing of an image by inverting it into a prior model such as GAN. With the strong prior bearing in a pre-trained generator, synthesizing novel views of a given portrait can be achieved

without any 3D or video data during training. Among them, methods using 3D-aware GANs [7,9,13,46,47] have shown better 3D consistency under pose variations. Our method is also based on 3D-aware GAN and largely improves the efficiency, reconstruction quality, as well as 3D consistency.

3. Approach

Given a monocular portrait image $\hat{I}$, we aim to synthesize its novel views at some arbitrary camera viewpoints by leveraging the prior knowledge of a pre-trained 3D-aware GAN, as shown in Fig. 2. To guarantee high-quality and 3D-consistent novel view synthesis, we adopt GRAM [13] as our underlying image generator and design an efficient version of it that requires much less computation and memory cost so as to incorporate it into our whole framework (Sec. 3.1). With the efficient GRAM, we first utilize a *general encoder-based GAN inversion* to reconstruct the coarse radiance manifolds from the input image (Sec. 3.2). We then introduce a *detail-specific reconstruction stage* to learn high-resolution detail manifolds that cannot be well captured by the coarse result, via our proposed novel detail manifolds reconstructor (Sec. 3.3). Multiple losses are enforced to regulate the predicted detail manifolds to ensure reasonable synthesized results at novel views, by leveraging the 3D priors derived from the coarse radiance manifolds (Sec. 3.4). We describe each part in details below.

3.1. Efficient Generative Radiance Manifolds

We start with a brief review of the original GRAM proposed in [13]. The core of GRAM is its underlying radiance manifolds representation, which regulates radiance field learning on a set of surface manifolds in the 3D space instead of predicting it in the whole volumetric space as done by [29]. The surface manifolds are defined as a set of iso-surfaces $\{\mathcal{S}_i\}$ in a 3D scalar field represented by a

light-weight MLP called the manifold predictor $\mathcal{M}$:

$$\mathcal{M} : \boldsymbol{x} \in \mathbb{R}^3 \rightarrow s \in \mathbb{R}, \ \mathcal{S}_i = \{\boldsymbol{x} | \mathcal{M}(\boldsymbol{x}) = l_i\}, \quad (1)$$

where $\{l_i\}$ are N predefined scalar levels. During image generation, only intersections $\{\boldsymbol{x}_i\}$ between a viewing ray $\boldsymbol{r}$ and the surface manifolds will be sent into an MLP-based radiance generator Φ for radiance prediction:

$$\Phi : (\boldsymbol{z}, \boldsymbol{x}_i) \in \mathbb{R}^{d_z} \times \mathbb{R}^3 \rightarrow (\boldsymbol{c}, \alpha) \in \mathbb{R}^4, \quad (2)$$

where $\boldsymbol{z}$ is a latent code determining the radiance, $\boldsymbol{c}$ is the color, and α is the occupancy. The final color of each ray can be computed via manifold rendering [13, 71]:

$$C(\boldsymbol{r}) = \sum_{i=1}^{N} \prod_{j<i} (1 - \alpha(\boldsymbol{x}_j)) \alpha(\boldsymbol{x}_i) \boldsymbol{c}(\boldsymbol{x}_i). \quad (3)$$

The high momery cost of GRAM lies in its MLP-based radiance generator Φ, which requires millions of forward steps to generate a single image. Inspired by the recent EG3D [9], we substitute the original radiance generator with a tri-plane generator [9] based on StyleGAN2 structure [26]. Its efficient coarse-to-fine structure helps to reduce memory and computation costs by a large margin. Given the new radiance generator, the color and occupancy for points on the surface manifolds can be obtained by first generating tri-plane features by a 2D CNN Ψ, and then conducting tri-plane sampling and sending the sampled features into a small MLP-based decoder m as done in [9]. Note that although we take the tri-plane generator from EG3D to improve efficiency, we do not use its 2D super-resolution module but keep strictly to the radiance manifolds representation. This helps us to maintain the strong 3D consistency brought by the manifold rendering. In addition, we calculate ray-manifold intersections at $1/4$ resolution of the final image to further speed up our image generation process (details in the *suppl. material*).

The efficient GRAM serves as a strong prior for generating realistic multiview images of virtual subjects. By combining it with our two-stage manifolds reconstruction method, we achieve high-quality novel view synthesis of real portraits, as described in the following sections.

3.2. General Inversion Stage

Given a pre-trained efficient GRAM following a typical 3D-aware GAN training paradigm [13], we first introduce an image inverter E_w that maps a given image to the latent space of the efficient GRAM, as shown in Fig. 2. Inspired by previous StyleGAN-based inversion methods [1, 52, 55], we invert the given image into a latent code $\boldsymbol{w}+ = [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_L]$ in $\mathcal{W}+$ space [1] of the tri-plane generator Ψ for a proper trade-off between inversion fidelity and pose editing quality, where L is the number of layers in Ψ's synthesis sub-network. We leverage
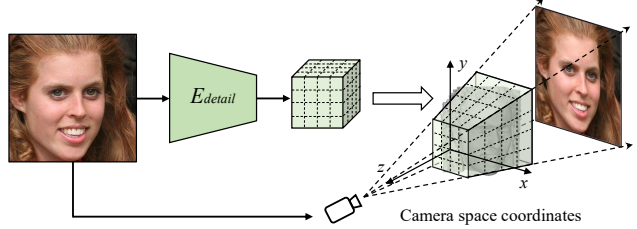

Figure 3. The detail encoder E_{detail} extracts a camera space feature voxel from an input image. The voxel corresponds to a quadrangular frustum in the 3D space bounded by the near and far planes of the camera and the outermost viewing rays.

the e4e encoder [52] as the backbone of E_w. Given $\boldsymbol{w}+$, we can obtain a coarse radiance manifolds $\Phi(\boldsymbol{w}+, \{\mathcal{S}_i\}) = m \circ \Psi(\boldsymbol{w}+, \{\mathcal{S}_i\})$ via Eq. (1) and (2), and further obtain a coarse inversion image I_w by rendering the radiance manifolds at input viewpoint $\hat{\boldsymbol{\theta}}$ via Eq. (3), where $\hat{\boldsymbol{\theta}}$ can be obtained by off-the-shelf 3D face reconstruction method [14].

We fixed the pre-trained efficient GRAM and learn the image inverter E_w following the training process of [52], except that we replaced the adversarial loss in [52] with a naive L2 loss between predicted $\boldsymbol{w}+$ and the average latent code of the $\mathcal{W}+$ space. To further improve the reconstruction fidelity, we fixed the trained E_w and finetuned the efficient GRAM via the pivot tuning strategy [38] using all training images. Details for the above training processes can be found in the *suppl. material*.

After training, the general inversion stage can already synthesize reasonable multiview images of the input, yet it cannot faithfully preserve the fine details, making the inverted result looks less like the original image (see Fig. 6). Therefore, we introduce a detail-specific stage for faithful detail reconstruction, as described below.

3.3. Detail-Specific Reconstruction Stage

The detail-specific stage aims to extract fine details from the input image that cannot be well described by the coarse radiance manifolds to improve the reconstruction fidelity. The intuition is to learn high-resolution details in 3D space so that their combination with the coarse radiance manifolds still remains strong 3D consistency under pose variations. To achieve this goal, we design a detail manifolds reconstructor consisting of two modules: a detail encoder E_{detail} that extracts low-resolution feature voxel from the input image, and a super-resolution module $\mathcal{U}$ to predict high-resolution detail manifolds from the low-resolution voxel.

Specifically, E_{detail} takes an image $\hat{I}$ as input[1] and predicts a camera space feature voxel as shown in Fig. 3:

$$E_{detail} : \hat{I} \in \mathbb{R}^{H \times W \times 3} \rightarrow V \in \mathbb{R}^{H_{lr} \times W_{lr} \times D_{lr} \times d_V}, \quad (4)$$

---

[1]In practice, we find that concatenating $\hat{I}$ with an extra difference map $\Delta = \hat{I} - I_w$ as input yields better reconstruction quality, where I_w is the inversion image obtained by the general inversion stage.

where d_V is the feature dimension. We implement E_{detail} as a 3D U-Net with skip connections to extract both global geometry structures as well as local fine textures from the input image. We refer the readers to the *suppl. material* for detailed network structure.

The feature voxel is defined in camera space instead of world space (*i.e.* space where tri-plane features of efficient GRAM are defined) as it is easier for E_{detail} to extract image-aligned features than to learn transformed world space features (see Sec. 4.3). Given the feature voxel V, we can obtain the corresponding feature $f^{lr} \in \mathbb{R}^{d_V}$ for a point $\boldsymbol{x} \in \mathbb{R}^3$ in the world space via:

$$f^{lr} = \text{grid\_sample}(V, \text{world2cam}(\boldsymbol{x})), \quad (5)$$

where grid_sample is a tri-linear interpolation function, and world2cam is a rigid transformation function between the world space and the camera space.

Nevertheless, since V is a low-resolution voxel, directly combining it with the feature manifolds obtained from the general inversion stage leads to a blurry inversion result, while predicting a high-resolution voxel (*e.g.* 256^3) instead causes unaffordable memory cost. Inspired by [60], we take the advantage of our radiance manifolds representation to obtain high-resolution detail manifolds from the low-resolution voxel via manifold super-resolution. Specifically, we first obtain low-resolution detail manifolds $f^{lr}(\{S_i\})$ by querying features from V via Eq. (5) for low-resolution points grid on the surface manifolds $\{S_i\}$. We then flatten each manifold to a low-resolution feature map F_i^{lr} and send it to the super-resolution module $\mathcal{U}$ to obtain a high-resolution feature map $F_i^{hr} = \mathcal{U}(F_i^{lr})$, where $\mathcal{U}$ is a simple 2D CNN of 4 convolution blocks and 2 bilinear upsampling blocks. Finally, we obtain the high-resolution detail manifolds $f^{hr}(\{S_i\})$ by re-projecting each flattened feature map F_i^{hr} to the surface manifolds. Since we conduct super-resolution for 3D space surface manifolds, 3D consistency across different views can be naturally maintained during this process. Note that although the manifold super-resolution strategy is proposed in [60], it does not leverage it in a reconstruction scenario but to generate random fine details. By contrast, we utilize it for faithful reconstruction of high-frequency details in the original image.

Given $f^{hr}(\{S_i\})$ from the detail-specific stage, we add it to the coarse feature manifolds $\Psi(\boldsymbol{w}+, \{S_i\})$ from the general inversion stage, and send each feature point on the manifolds to the MLP-based decoder m to obtain the final radiance manifolds, as shown in Fig. 2. The final inversion image I can then be obtained similarly via manifold rendering at input view $\hat{\boldsymbol{\theta}}$. Novel views can also be easily generated given an arbitrary camera pose $\boldsymbol{\theta}$ during rendering.
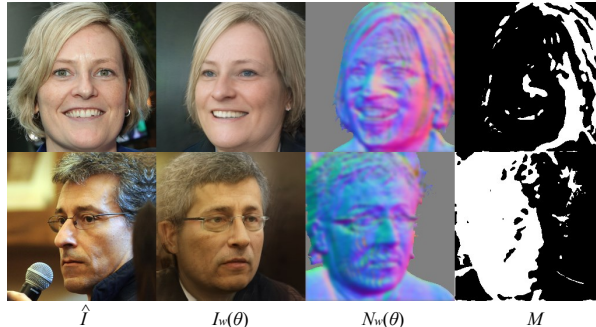

Figure 4. Visualization of the novel view regularization.

3.4. Detail Manifolds Learning

We fix the image inverter E_w and the efficient GRAM from the general inversion stage, and learn the detail manifolds reconstructor with the following losses.

Image reconstruction loss. A multi-level reconstruction loss is applied between the final inversion image I and the input image $\hat{I}$:

$$\mathcal{L}_r = ||I - \hat{I}||^2 + \text{LPIPS}(I, \hat{I}) + (1 - \langle f_{id}(I), f_{id}(\hat{I}) \rangle), \quad (6)$$

where $\text{LPIPS}(\cdot, \cdot)$ is the perceptual loss defined by [68], and f_{id} is a pre-trained face recognition network [11].

Novel view regularization. The reconstruction loss guarantees a faithful inversion result at the input viewpoint, yet artifacts can still occur when rendering the radiance manifolds at other views. We, therefore, design a regularization term to ensure reasonable novel view synthesis results:

$$\mathcal{L}_{nv} = \text{LPIPS}(M \odot I(\boldsymbol{\theta}), M \odot I_w(\boldsymbol{\theta})), \quad (7)$$

where $I(\boldsymbol{\theta})$ and $I_w(\boldsymbol{\theta})$ are final and coarse inversion image rendered at novel view $\boldsymbol{\theta}$ respectively, $\odot$ is element-wise multiplication, and M is a binary mask:

$$M(u, v) = \mathbb{I}(-\boldsymbol{r}(\hat{\boldsymbol{\theta}}) \cdot N_w(\boldsymbol{\theta})(u, v) < \tau). \quad (8)$$

Here (u, v) is the image space coordinate, $\mathbb{I}$ is the indicator function, $\boldsymbol{r}(\hat{\boldsymbol{\theta}})$ is the camera lookat direction of the input image $\hat{I}$, $N_w(\boldsymbol{\theta})$ is the surface normal map of $I_w(\boldsymbol{\theta})$, and τ is a scalar threshold. The intuition behind this regularization is that the details for regions unobserved in the input image should stay close to the coarse inversion result at novel views, as the coarse inversion image is more reasonable at new views due to leveraging the priors from the pre-trained GRAM. The normal map $N_w(\boldsymbol{\theta})$ in Eq. (8) can be effectively calculated via the following equation:

$$N_w(\boldsymbol{\theta})(u, v) = -\frac{1}{\eta} \sum_{i=1}^{N} T(\alpha(\boldsymbol{x}_i)) \alpha(\boldsymbol{x}_i) \frac{\partial \alpha(\boldsymbol{x}_i)}{\partial \boldsymbol{x}_i}, \quad (9)$$

where $\boldsymbol{x}_i$ are intersections along the ray that (u, v) corresponds to, $T(\alpha(\boldsymbol{x}_i)) = \prod_{j<i}(1 - \alpha(\boldsymbol{x}_j))$ is the accumulated transparency, and η is a normalizing scalar. The partial

gradient $\partial\alpha(\boldsymbol{x}_i)/\partial\boldsymbol{x}_i$ can be easily computed via backpropagating the MLP-based decoder m. Visualizations of the normal map $N_w(\boldsymbol{\theta})$ and the binary mask M are in Fig. 4.

Depth regularization. We further enforce a depth regularization to the HR detail manifolds $f^{hr}(\{S_i\})$ to ensure that details are predicted near the geometry surface:

$$\mathcal{L}_{depth} = \begin{cases} \lambda\|f^{hr}(\boldsymbol{x}_i)\|^2 & |z(\boldsymbol{x}_i) - z_{surf}| > \epsilon \\ 0 & |z(\boldsymbol{x}_i) - z_{surf}| \leq \epsilon \end{cases}, \quad (10)$$

where $\boldsymbol{x}_i$ are intersections along the viewing rays at input viewpoint $\hat{\boldsymbol{\theta}}$, $z(\boldsymbol{x}_i)$ is the depth of $\boldsymbol{x}_i$, $z_{surf} = \sum_{i=1}^{N} T(\alpha(\boldsymbol{x}_i))\alpha(\boldsymbol{x}_i)z(\boldsymbol{x}_i)$ is the depth of the approximated surface, and ϵ is a threshold. This regularization ensures correct parallax for the learned details (see Sec. 4.3).

4. Experiments

Implementation details. We train our method on the FFHQ [25] dataset at 256^2 resolution and test it on the CelebA-HQ [23] dataset. All images are pre-processed following the procedure in [13]. The camera pose of input images is estimated by the face reconstruction method of [14]. We train our models on 4 Tesla V100 GPUs with 32GB memory. The whole training process takes around 6 days, where training the efficient GRAM takes 2 days, training the image inverter and finetuning the efficient GRAM takes 2 days and 1 day respectively, and training the detail manifolds reconstructor takes 1 day. More in the *suppl. material*.

4.1. Novel View Synthesis Results

Figure 1 shows the novel view synthesis results of our method given different portrait images. Our method well preserves fine details (*e.g.* hair bangs, wrinkles, moles) of the input images and produces their 3D consistent novel views. The whole inversion and novel view synthesis process runs at 3 FPS on a V100 GPU without specialized acceleration, which largely improves the efficiency upon previous optimization-based 3D-aware GAN inversions. With the manifold caching technique in [13], we can further increase the free view rendering speed to 180FPS. More examples and video results are in the *suppl. material*.

4.2. Comparison with Prior Arts

Comparison with GRAM. We first compare our efficient version of GRAM with the original one [13]. We measure the image generation quality by the Fréchet Inception Distances (FID) [21] between 20K randomly generated images and 20K sampled real images. The 3D consistency is measured by the reconstruction quality of NeuS [54] (*i.e.* PSNR_{mv} and SSIM_{mv}) on multiview images of 50 generated instances following [60]. As shown in Tab. 1, our efficient GRAM largely reduces the memory cost and increases

Table 1. Comparison between efficient GRAM and GRAM [13]. *: Inference on a Tesla V100 GPU with a batchsize of 1.

Methods	Memory* ↓	FPS* ↑	FID ↓	PSNR_{mv} ↑	SSIM_{mv} ↑
EG3D	2.8G	20	6.02	34.0	0.928
GRAM	12G	2	15.0	**38.0**	0.966
Ours	**3.3G**	**14**	**14.2**	37.6	**0.969**

the inference speed upon the original one without sacrificing image generation quality or 3D consistency, by introducing the StyleGAN2-based radiance generator and the efficient intersection calculation strategy. This improvement enables our following GRAMinverter method, otherwise, it is difficult, if not impossible, to leverage the memory-consuming GRAM for encoder-based GAN inversion. We also list the performance of the state-of-the-art EG3D [9] as a reference. Although EG3D has better image quality, it sacrifices the 3D consistency which we argue is a key factor for 3D-aware generation.

Comparison with pose editing methods. We compare with existing methods that achieve 3D pose editing of a given portrait via single forward pass, including 2D GAN inversion-based methods: e4e [52]+InterFaceGAN [41], HFGI [55]+InterFaceGAN, and HFGI+StyleHEAT [65]; 3D-aware GAN inversion-based methods: pix2NeRF [7] and IDE-3D [46]; and face reenactment methods: PIRenderer [36] and Face-vid2vid [56].

We first evaluate the inversion fidelity among GAN inversion-based methods (Tab. 2). We report PSNR, SSIM, LPIPS, identity similarity (*i.e.* ID) measured by cosine distance of face recognition features [53], and FID. All metrics are calculated between the first 1K images of CelebA-HQ and their corresponding inverted results. Since different methods may generate images of different resolutions and alignments, we pre-process all results following [13] and resize them to 256×256 for a fair comparison. As shown, our method significantly outperforms other 3D-aware GAN inversion methods across all metrics. We also exceed the StyleGAN2-based inversion method e4e and achieve comparable results with the state-of-the-art method HFGI.

We further compare our method with other approaches on pose editing of portrait images. We generate novel views (see the *suppl. material*) of the 1K test images using different methods and evaluate their identity similarity and FID to the original input in Tab. 2. Higher ID_{nv} and lower FID_{nv} indicate that a method can better keep the identity and image quality while changing the camera pose. Our method yields the best result among all competitors. We also surpass PIRenderer and Face-vid2vid which require video data for training, while ours is merely trained on monocular in-the-wild images. Figure 5 shows a visual comparison.

Finally, we measure the 3D consistency of all methods during continuous variation of the camera viewpoint. Fol-

Table 2. Quantitative comparison with existing portrait editing methods. See the text for details.

| Methods | Inversion fidelity | | | | | Novel view quality | | 3D consistency | |
	PSNR ↑	SSIM ↑	LPIPS ↓	ID ↑	FID ↓	ID$_{nv}$ ↑	FID$_{nv}$ ↓	PSNR$_{mv}$ ↑	SSIM$_{mv}$ ↑
PIRenderer [36]	–	–	–	–	–	0.476	42.64	36.72	0.958
Face-vid2vid [56]	–	–	–	–	–	0.416	41.76	36.10	0.942
e4e [52] + InterfaceGAN [41]	19.23	0.451	0.213	0.706	35.92	0.489	38.04	34.29	0.909
HFGI [55] + InterfaceGAN [41]	22.30	0.579	0.135	0.827	26.41	0.516	45.23	33.99	0.917
HFGI [55] + StyleHEAT [65]	22.30	0.579	0.135	0.827	26.41	0.457	58.33	35.93	0.951
pix2NeRF [7]	16.95	0.394	0.452	0.466	108.3	0.378	115.6	50.66	0.997
IDE-3D (encoder) [46]	16.73	0.382	0.290	0.393	51.51	0.324	47.56	37.57	0.950
Ours	21.51	0.650	0.127	0.936	28.17	0.635	36.02	39.53	0.974
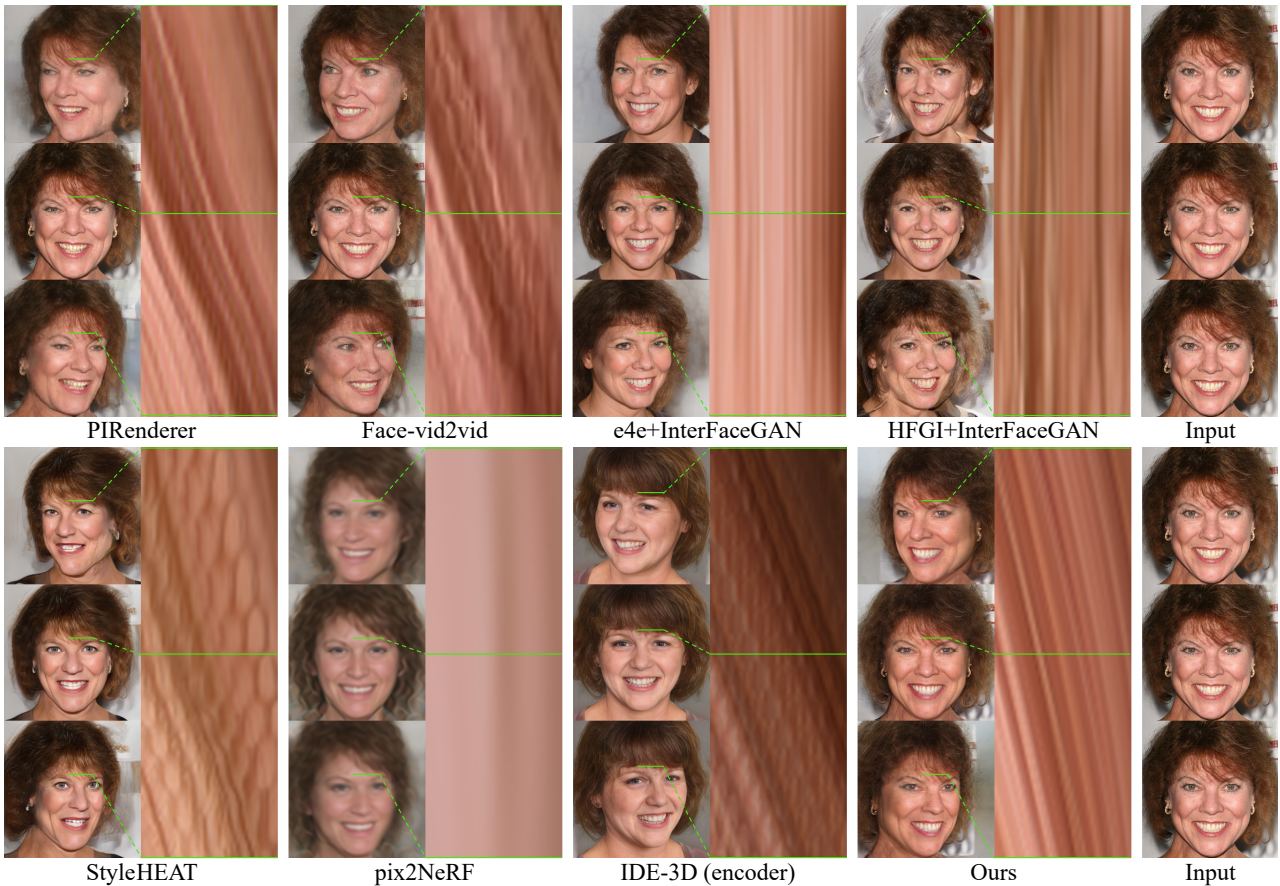
Figure 5. Pose editing comparison. Texture images with smoothly tilted strips indicate better 3D consistency. **Best viewed with zoom-in.**

lowing [60], for each method, we generate 30 images under different views for 50 test instances in CelebA-HQ, and measure the multiview reconstruction quality of NeuS on them (*i.e.* PSNR$_{mv}$ and SSIM$_{mv}$). In theory, better 3D consistency across different views would reduce the learning difficulty of NeuS, thus leading to higher PSNR and SSIM. Table 2 shows that our method has the second best 3D consistency among all methods, while the best one (*i.e.* pix2NeRF) generates over-smooth images of low quality as shown in Fig. 5 and indicated by the high FID score in Tab. 2. Our method outperforms IDE-3D in that it utilizes

a 2D super-resolution module in its 3D-aware GAN which lowers the 3D consistency to some extent. Nevertheless, all 3D-aware GAN-based methods yield better 3D consistency compared to other 2D methods, indicating the importance of 3D-aware GAN for pose editing of images. A further comparison with the full pipeline of IDE-3D which includes an extra optimization step is in the *suppl. material*.

Figure 5 further shows the visual comparison of 3D consistency, where we draw the stacked texture image of a fixed horizontal line segment during continuous camera movement following [60]. Methods with strong 3D consistency

Figure 6. Visual comparison between different alternatives. Our final solution yields the best result. **Best viewed with zoom-in.**

Table 3. Ablation study of our proposed framework.

Methods	PSNR ↑	LPIPS ↓	ID ↑	FID$_{nv}$ ↓
General	17.68	0.265	0.648	44.43
General - pretrain	17.45	0.280	0.472	52.12
General + finetune	18.00	0.254	0.678	43.46
Detail (Ours)	21.51	0.127	0.936	36.02
Detail - world2cam	19.96	0.171	0.908	38.74
Detail - superres	21.06	0.165	0.926	38.45
Detail - $\mathcal{L}_{nv}$	22.32	0.106	0.949	36.49
Detail - mask	19.08	0.211	0.840	43.25
Detail - $\mathcal{L}_{depth}$	23.68	0.094	0.960	35.60

will result in texture images with smoothly tilted strips, while methods with low 3D consistency produce twisted textures (*i.e.* geometry distortions and texture flickering issues) or vertical lines (*i.e.* texture sticking issues). Our method clearly produces a more reasonable texture image compared to the others. See the *suppl. material* and the accompanying video for more results.

4.3. Ablation Study

We conduct an ablation study to validate the efficacy of our proposed framework and report the results in Tab. 3 and Fig. 6. All metrics are calculated similarly as in Sec. 4.2.

Inversion stage. Table 3 shows the performance of different stages. *General* stands for the general inversion stage without finetuned generator. *General - pretrain* denotes learning the efficient GRAM with E_w together instead of pre-training it via the 3D-aware GAN framework. *General + finetune* denotes finetuning the pre-trained efficient GRAM as described in Sec. 3.2, and *Detail* denotes our final approach with detail-specific reconstruction. As shown,

the general stage alone cannot produce faithful reconstruction result, whether the efficient GRAM is further finetuned or not. By contrast, introducing the detail-specific reconstruction significantly improves the inversion fidelity upon the previous stage without sacrificing novel view quality. Learning the efficient GRAM together with the encoder without pre-training leads to a significant performance drop, indicating the importance of leveraging the prior knowledge of a pre-trained 3D-aware GAN.

Network architecture. We ablate the architecture of the detail manifolds reconstructor. As shown in Tab. 3 and Fig. 6, learning the low-resolution detail voxel in world space instead of in camera space (- *world2cam*) harms the reconstruction fidelity. And removing the super-resolution module for high-resolution manifold prediction (- *superres*) leads to blurry inversion results.

Regularization. We further validate our proposed regularization for detail manifolds learning. As shown in Tab. 3 and Fig. 6, removing the novel view regularization (- $\mathcal{L}_{nv}$) causes obvious artifacts at new views and leads to the increase of FID$_{nv}$, though it improves the reconstruction quality at the input viewpoint. Simply enforcing $\mathcal{L}_{nv}$ without the normal-aware mask (- *mask*) damages fine texture preservation at visible regions. Finally, although learning without the depth regularization (- $\mathcal{L}_{depth}$) results in better metrics, we find that it cannot well preserve certain fine details at novel views due to incorrect parallax brought by the depth error (*e.g.* mole in Fig. 6). We conjecture that such dynamic artifact can hardly be captured by the current feature extractor [49] for FID computation.

5. Conclusions

We presented GRAMinverter, a novel approach for high-fidelity and 3D-consistent portrait synthesis from monocular images via single forward pass. The core idea is to learn a detail manifolds reconstructor to predict 3D-consistent fine details on the radiance manifolds from a input image, and combine them with the coarse radiance manifolds obtained via an encoder-based inversion of the pre-trained GRAM. Extensive experiments have demonstrated our superior results over previous works. We believe our method paves a new way for efficient 3D-aware portrait creation.

Limitations and future works. Our GRAMinverter has several limitations. Based on the radiance manifold representation, it produces layered artifacts at large viewing angles. It cannot well handle occlusions of hands and other accessories. Its performance is also affected by the training data and may produce inferior results for out-of-distribution input. Besides, it does not support editing of attributes beyond camera viewpoints as done in previous 2D GAN inversions. Better 3D representations and inversion strategies should be further explored to alleviate these problems.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 1, 2, 4

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8296–8305, 2020. 2

[3] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics*, 40(3):1–21, 2021. 2, 3

[4] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 2

[5] Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third time's the charm? image and video editing with stylegan3. *arXiv preprint arXiv:2201.13433*, 2022. 2

[6] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4502–4511, 2019. 2

[7] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2nerf: Unsupervised conditional p-gan for single image to neural radiance fields translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3981–3990, 2022. 2, 3, 6, 7

[8] Menglei Chai, Linjie Luo, Kalyan Sunkavalli, Nathan Carr, Sunil Hadap, and Kun Zhou. High-quality hair modeling from a single portrait photo. *ACM Transactions on Graphics (TOG)*, 34(6):1–10, 2015. 3

[9] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3, 4, 6

[10] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5799–5809, 2021. 2, 3

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 5

[12] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3D imitative-contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5154–5163, 2020. 1, 3

[13] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *IEEE Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 4, 6

[14] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 4, 6

[15] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14398–14407, 2021. 3

[16] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. *arXiv preprint arXiv:2207.07621*, 2022. 3

[17] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In *Advances In Neural Information Processing Systems*, 2022. 2

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 1

[19] 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. 2

[20] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 1, 3

[21] 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. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 6

[22] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 3

[23] 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. 6

[24] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 2

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 2, 6

[26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 8110–8119, 2020. 1, 2, 4

[27] Connor Z Lin, David B Lindell, Eric R Chan, and Gordon Wetzstein. 3d gan inversion for controllable portrait image animation. *arXiv preprint arXiv:2203.13441*, 2022. 2, 3

[28] BR Mallikarjun, Ayush Tewari, Abdallah Dib, Tim Weyrich, Bernd Bickel, Hans Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Louis Chevallier, Mohamed A Elgharib, et al. Photoapp: Photorealistic appearance editing of head portraits. *ACM Transactions on Graphics*, 40(4), 2021. 3

[29] 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. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 2, 3

[30] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3D representations from natural images. In *IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 2

[31] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2

[32] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019. 3

[33] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *IEEE/CVF International Conference on Computer Vision*, 2021. 2

[34] 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 *IEEE/CVF International Conference on Computer Vision*, 2022. 2

[35] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. *arXiv preprint arXiv:2011.00844*, 2020. 3

[36] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021. 3, 6, 7

[37] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 2

[38] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021. 1, 2, 4

[39] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems*, 2020. 2

[40] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *arXiv preprint arXiv:2206.07695*, 2022. 2

[41] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020. 1, 3, 6, 7

[42] Yichun Shi, Divyansh Aggarwal, and Anil K Jain. Lifting 2d stylegan for 3d-aware face generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6258–6266, 2021. 2

[43] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 3

[44] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[45] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. *arXiv preprint arXiv:2206.10535*, 2022. 2

[46] 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 preprint arXiv:2205.15517*, 2022. 2, 3, 6, 7

[47] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2022. 2, 3

[48] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3d shape learning from natural images. *arXiv preprint arXiv:1910.00287*, 2019. 2

[49] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 8

[50] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 1, 3

[51] Ayush Tewari, Xingang Pan, Ohad Fried, Maneesh Agrawala, Christian Theobalt, et al. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1516–1525, 2022. 3

[52] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 1, 2, 4, 6, 7

[53] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface:

Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 6

[54] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2, 6

[55] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022. 2, 3, 4, 6, 7

[56] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 3, 6, 7

[57] Ziyu Wang, Yu Deng, Jiaolong Yang, Jingyi Yu, and Xin Tong. Generative deformable radiance fields for disentangled image synthesis of topology-varying objects. *arXiv preprint arXiv:2209.04183*, 2022. 3

[58] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018. 3

[59] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2020. 3

[60] 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. 2, 5, 6, 7

[61] Sicheng Xu, Jiaolong Yang, Dong Chen, Fang Wen, Yu Deng, Yunde Jia, and Xin Tong. Deep 3d portrait from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7710–7720, 2020. 3

[62] Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[63] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. Feature-style encoder for style-based gan inversion. *arXiv e-prints*, pages arXiv–2202, 2022. 2, 3

[64] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 2

[65] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pretrained stylegan. *arXiv preprint arXiv:2203.04036*, 2022. 6, 7

[66] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020. 3

[67] Jichao Zhang, Aliaksandr Siarohin, Yahui Liu, Hao Tang, Nicu Sebe, and Wei Wang. Training and tuning generative neural radiance fields for attribute-conditional 3d-aware face generation. *arXiv preprint arXiv:2208.12550*, 2022. 2

[68] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5

[69] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. *arXiv preprint arXiv:2207.10642*, 2022. 2

[70] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 3

[71] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4):1–12, 2018. 4

[72] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020. 2

[73] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016. 2

[74] Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020. 2

[75] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 3

[76] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 787–796, 2015. 3