Injecting 3D Perception of Controllable NeRF-GAN into StyleGAN for Editable Portrait Image Synthesis

Jeong-gi Kwak[†] Yuanming Li[†] Dongsik Yoon[†] Donghyeon Kim[†]
David Han[‡] Hanseok Ko[†]

[†] Korea University [‡]Drexel Univiersity

Abstract. Over the years, 2D GANs have achieved great successes in photorealistic portrait generation. However, they lack 3D understanding in the generation process, thus they suffer from multi-view inconsistency problem. To alleviate the issue, many 3D-aware GANs have been proposed and shown notable results, but 3D GANs struggle with editing semantic attributes. The controllability and interpretability of 3D GANs have not been much explored. In this work, we propose two solutions to overcome these weaknesses of 2D GANs and 3D-aware GANs. We first introduce a novel 3D-aware GAN, SURF-GAN, which is capable of discovering semantic attributes during training and controlling them in an unsupervised manner. After that, we inject the prior of SURF-GAN into StyleGAN to obtain a high-fidelity 3D-controllable generator. Unlike existing latent-based methods allowing implicit pose control, the proposed 3D-controllable StyleGAN enables explicit pose control over portrait generation. This distillation allows direct compatibility between 3D control and many StyleGAN-based techniques (e.g., inversion and stylization), and also brings an advantage in terms of computational resources. Our codes are available at https://github.com/jgkwak95/SURF-GAN.

Keywords: 3D-aware portrait generation, pose-disentangled GAN, facial image editing, novel view synthesis, latent manipulation

1 Introduction

Since the advent of Generative Adversarial Networks (GANs) [15], remarkable progress has been made in the field of photorealistic image generation. The quality and diversity of images generated by 2D GANs have been improved considerably and recent models [25,6,28,29,27] can produce high resolution images at a level that humans cannot distinguish. Despite the expressiveness of 2D GANs, they lack 3D understanding, in that the underlying 3D geometry of an object is ignored in the generation process. As a result, they suffer the problem of multi-view inconsistency. To overcome the issue, many researchers have studied 3D controllable image synthesis and it has become one of the mainstream research

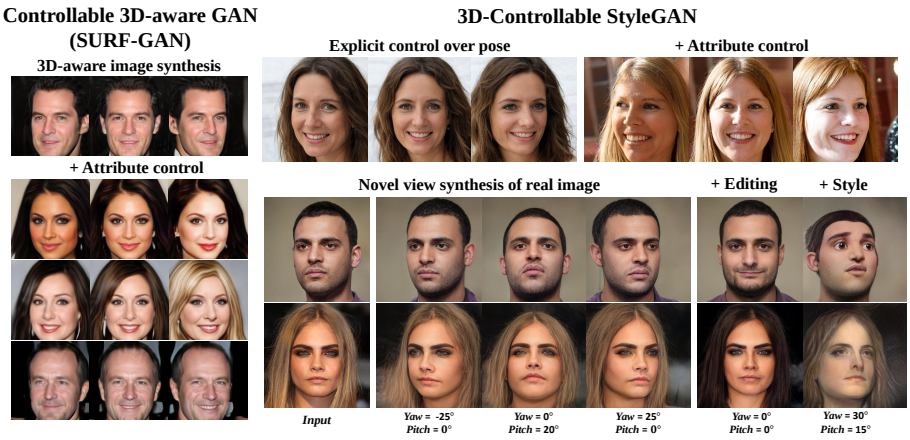

Fig. 1: (Left): Proposed novel 3D-aware GAN (SURF-GAN) which is capable of attribute-controllable generation as well as 3D-aware synthesis. (Right): 3D-controllable StyleGAN obtained by distilling the prior of SURF-GAN into 2D GAN.

in the community. There have been several attempts to learn 3D pose information with 2D GAN by disentangling pose in the latent space, but they require auxiliary 3D supervision such as synthetic face dataset [30] or 3DMM [12,58,57]. In addition, a few unsupervised approaches have been proposed by adopting implicit 3D feature [37,38] or differentiable renderer [52,42] in generation. However, these methods have struggled with multi-view consistency and photorealism.

Since the introduction of neural radiance fields (NeRF) by Mildenhall et al. [36] which has achieved notable success in novel view synthesis, a new paradigm has emerged in 3D-aware generation, called 3D-aware GAN. Several researchers have proposed 3D-aware generative frameworks [49,39,8,66,16,13,60] by leveraging NeRF as a 3D representation in GAN generator. NeRF-GANs learn 3D geometry from unlabelled images yet allow accurate and explicit control of 3D camera based on a volume rendering. Despite the obvious advantages, 3D GANs based on a pure NeRF network require tremendous computational resources and generate blurry images. Very recently, several approaches have alleviated the problems and have shown photorealistic output with high resolution by incorporating rear-end 2D networks [66,16,7,13,60]. However 3D GANs have difficulty with attribute-controllable generation or real image editing because their latent space has been rarely investigated for interpretable generation (Fig. 2).

In summary, these two distinct approaches have strengths and weaknesses that are complementary: 3D-aware GAN can generate novel poses but it has trouble with disentangling and manipulating attributes; 2D GAN is capable of controlling attributes but it struggles with 3D controllability. In this work, we propose novel solutions to overcome each weakness of 2D GANs and 3D GANs.

First, we propose a novel 3D-aware GAN, i.e., SURF-GAN, which can discover semantic attributes by learning layer-wise **SU**bspace in INR-based Ne**RF**

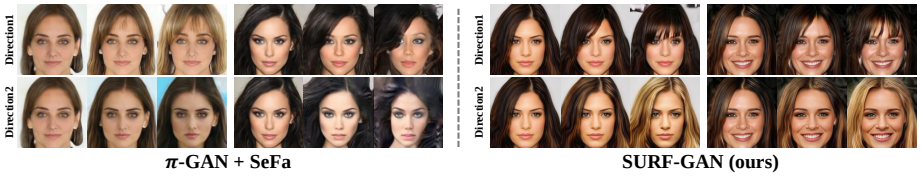

Fig. 2: Results of attribute manipulation when applying SeFa [51] utilized for 2D GANs to a 3D-aware GAN [8] (left). The captured attributes are entangled or not meaningful. In contrast, SURF-GAN can captures disentangled and semantic attributes (right).

network in an unsupervised manner. The discovered semantic vectors can be controlled by corresponding parameters, thus this property allows us to manipulate semantic attributes (e.g., gender, hair color, etc.) as well as explicit pose.

With the proposed SURF-GAN, we take one more step to transform StyleGAN into a 3D-controllable generator. We inject the prior of 3D-aware SURF-GAN into the expressive and disentangled latent space of 2D StyleGAN. Unlike the previous methods [50,17,51] that allows implicit pose control, we make StyleGAN enable explicit control over pose. It means that the generator is capable of synthesizing accurate images based on a conditioned target view. By utilizing SURF-GAN which consists of pure NeRF layers as a generator of pseudo multi-view images, the transformed StyleGAN can learn elaborate control over 3D camera pose with latent manipulation. To this end, we proposed a method to find several orthogonal directions (not a single) related to the same pose attribute, and explicit control over the pose is accomplished by a combination of these directions. With a GAN inversion encoder, 3D controllable StyleGAN can be extended to the task of novel pose synthesis from a real image.

In addition to 3D perception, we also inject the controllability about semantic attributes that SURF-GAN finds. We can find more pose-robust latent path in the latent space of StyleGAN because SURF-GAN can manipulate a specific semantic while keeping view direction unchanged. Moreover, it allows further applications related to StyleGAN family, e.g., 3D control over stylized images generated by fine-tuned StyleGAN. It is notable that our approach neither requires 3D supervision nor exploits auxiliary off-the-shelf 3D models (e.g., 3DMM or pose detector) in both training and inference because SURF-GAN learns 3D geometry from unlabelled 2D images from scratch.

In summary, our contributions are as follow:

- We propose a novel 3D-aware GAN, called SURF-GAN, which can discover controllable semantic attributes in an unsupervised manner.
- By injecting editing directions from the low-resolution 3D-aware GAN into the high-resolution 2D StyleGAN, we achieve a 3D controllable generator which is capable of explicit control over pose and 3D consistent editing.
- Our method is directly compatible with various well-studied 2D StyleGAN-based techniques such as inversion, editing or stylization.

2 Related Work

Pose-disentangled GANs. The remarkable advances have been achieved in photorealism by state-of-the-art GAN models [25,6,28,29,27]. However, pose control by image generators has been limited due to a lack of 3D understanding in the synthesizing process. Thereby, several works have attempted to disentangle the pose information from other attributes in 2D GANs. The disentanglement has been achieved by leveraging supervision such as 3DMM [12,58,57,62], landmark [21], synthetic images from 3D engine [30] or pose detector [53]. A few unsupervised approaches without 3D supervision [37,38] have been proposed by disentangling pose with implicit 3D feature projection, but they allow only implicit 3D control and show blurry results. Recently, a few methods [52,42] have incorporated a pre-trained StyleGAN with a differentiable renderer, but they struggle with photorealism, high-resolution [42] and real image editing [52].

Interpretabilty and controllabiltiy of GAN. The well-trained 2D GANs, such as StyleGAN [28,29] have shown capable of disentangling the latent space. Recent works [50,17,3,40,51,61,43] have demonstrated semantic manipulation, especially for facial attributes, by analyzing the manifold and finding meaningful direction or mapping. Combining with GAN inversion [1,67,2,47,59,48,4,5], the applications of 2D GANs have been extended to real image editing. Alternatively, there have been studies [10,24,32,22] that discover and disentangle latent embeddings into interpretable dimensions during training of the generator. EigenGAN [19] that inspired our approach has demonstrated interpretable latent dimensions by designing layer-wise subspace embedding. However, both types of methods support implicit control over the discovered semantics. In the case of a pose that can be defined with camera parameters, these methods struggle to synthesize explicit novel view elaborately. Of course, the implicit methods can eventually create the desired pose through manual and iterative adjustment, but this is not an ideal situation. We can obtain a frontalized image automatically with some latent-based methods [50,31,47], but not for arbitrary target pose. Recently, Chen et al. [9] have introduced a generator allowing explicit control over pose, but it requires 3D mesh for pre-training process.

3D-aware GANs. Beyond the disentanglement of pose information, many efforts have been made to obtain 3D-awareness in generation. Earlier methods have adopted several explicit 3D representations in 2D image generation such as voxel [68,35,20,14] or mesh [33,56]. However, they suffer from a lack of visual quality and limited resolution. Recently, approaches [49,8,39,16,66,7,13,41,60] based on neural fields have made significant progress in photorealism and 3D consistency. Nevertheless, these 3D-aware GANs have weakness in finding and editing semantic attribute because their latent space has been rarely investigated. Very recently, Sun et al. [55] have proposed an editable NeRF-GAN, but it does not handle diverse semantic attributes and requires semantic maps as supervision. In addition, 3D GANs struggle with novel pose generation of real image

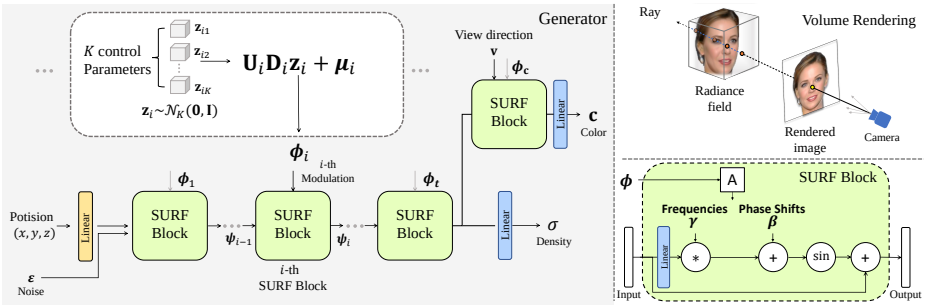

Fig. 3: Overview of SURF-GAN generator. Interpretable dimensions are caputured in layers with sub-modulation vectors. As like INR 3D-aware GANs, it takes position and view direction as input and predicts view dependent color ($\mathbf{c}$) and its density (σ).

despite their capability of multi-view consistency. Recently proposed EG3D [7] has shown experiments of novel view synthesis and presented outstanding results, but it requires iterative optimization for latent code and fine-tuning of the generator [48] for each target image.

3 Proposed method

In this section, we describe our method, by first introducing SURF-GAN in detail and then by explaining a method to inject the prior of 3D SURF-GAN into 2D StyleGAN. Note that the word "StyleGAN" denotes StyleGAN2 [29].

3.1 Towards controllable NeRF-GAN

Preliminaries: NeRF-GANs. Existing 2D GANs (e.g., StyleGAN [28,29]) synthesize output image directly with sampled latent vector. However, NeRF-GANs [8,39,16,66] generate a radiance field [36] before rendering 2D image. Given a position $\mathbf{x} \in \mathbb{R}^3$ and a viewing direction $\mathbf{v} \in \mathbb{S}^2$, it predicts a volume density $\sigma(\mathbf{x}) \in \mathbb{R}_+$ and the view-dependent RGB color $\mathbf{c}(\mathbf{x}, \mathbf{v}) \in \mathbb{R}^3$ of the input point. The points are sampled from rays of camera, and then an image is rendered into 2D grid with a classic volume rendering technique [23]. To produce diverse images, existing NeRF-GAN methods adopt StyleGAN-like modulation, where some components in the implicit neural network, e.g., intermediate features [8,66] or weight matrices [16] are modulated by sampled noise passing though a mapping network. Thereby, NeRF-GAN can control the pose by manipulating viewing direction $\mathbf{v}$ and change identity by injecting different noise vector. Nevertheless, it is ambiguous how to interpret the latent space and how to disentangle semantic attributes of NeRF-GAN for controllable image generation.

Learning layer-wise subspace in NeRF network. Inspired by EigenGAN [19], we adopt a different strategy from the existing methods [8,66] those modulation is obtained by the mapping network consisting of several MLPs. EigenGAN

learns interpretable subspaces in layers of its generator during training. However, EigenGAN is typical 2D convolution-based GAN framework, thus its concept is inapplicable to INR based NeRF-GAN. Therefore, we propose a novel framework (i.e., SURF-GAN), which captures the disentangled attributes in layers of NeRF network. Fig. 3 shows the overview of SURF-GAN. The generator consists of $t+1$ SURF blocks (t for shared layers and one for color layer). Following π-GAN, SURF block adopts the feature-wise linear modulation (FiLM) [44] to transform the intermediate features with frequencies $\boldsymbol{\gamma}_i$ and phase shifts $\boldsymbol{\beta}_i$, and followed SIREN activation [54]. SURF block in i^{th} layer is formulated as

$$\boldsymbol{\psi}_i = \text{SURF}_i\left(\boldsymbol{\psi}_{i-1}, \boldsymbol{\phi}_i\right) = \sin\left(\boldsymbol{\gamma}_i \cdot \left(\mathbf{W}_i \boldsymbol{\psi}_{i-1} + \mathbf{b}_i\right) + \boldsymbol{\beta}_i\right) + \boldsymbol{\psi}_{i-1}, \tag{1}$$

where $\boldsymbol{\psi}_{i-1}$ and $\boldsymbol{\phi}_i$ denote input feature and modulation of i^{th} layer respectively. $\mathbf{W}_i$ and $\mathbf{b}_i$ represent the weight matrix and followed bias. Unlike other NeRF-GANs, we add skip connection [18] to prevent drastic change of modulation vectors in training. In the model, a subspace embedded in each layer determines the modulation. Each subspace has orthogonal basis and it can be updated during training. The basis are learned to capture semantic modulation. Concretely, in the case of i^{th} layer, a specific subspace determines the modulation of i^{th} layer of NeRF network. It consists of learnable matrices, orthonormal basis $\mathbf{U}_i = [\mathbf{u}_{i1}, \ldots, \mathbf{u}_{iK}]$ and a diagonal matrix $\mathbf{D}_i = \text{diag}\left(d_{i1}, \ldots, d_{iK}\right)$. Each column of $\mathbf{U}_i$ plays a role of sub-modulation and it is updated to discover a meaningful direction that results in semantic change in image space. $d_{i1}, \ldots, d_{iK}$ serve as scaling factors of corresponding basis vectors $\mathbf{u}_{i1}, \ldots, \mathbf{u}_{iK}$. The latent $\mathbf{z}_i \in \mathbb{R}^K$ is set of K scalar control parameters, i.e.,

$$\mathbf{z}_i = \left\{z_{ij} \in \mathbb{R} \mid z_{ij} \sim \mathcal{N}(0,1), j = 1, \ldots, K\right\}, \tag{2}$$

where z_{ij} is a coefficient of sub-modulation $d_{ij}\mathbf{u}_{ij}$. Hence, the modulation of i^{th} layer ϕ_i is decided by weighted summation of K sub-modulations with $\mathbf{z}_i$, i.e,

$$\boldsymbol{\phi}_i = \mathbf{U}_i \mathbf{D}_i \mathbf{z}_i + \boldsymbol{\mu}_i = \sum_{j=1}^{K} z_{ij} d_{ij} \mathbf{u}_{ij} + \boldsymbol{\mu}_i, \tag{3}$$

where the marginal vector $\boldsymbol{\mu}_i$ is employed to capture shifting bias. Finally, a simple affine transformation is applied to $\boldsymbol{\phi}_i$ for matching dimension and obtaining frequency $\boldsymbol{\gamma}_i$ and phase shift $\boldsymbol{\beta}_i$. At training phase, SURF-GAN layers learn variations of meaningful modulation controlled by randomly sampled $\mathbf{z}$. Additionally, an input noise ϵ is also injected to capture the rest variations missed by the layers. To improve the disentanglement of attributes and to prevent the basis fall into a trivial solution, we adopt the regularization loss to guarantee the column vectors of $\mathbf{U}_i$ to be orthogonal following EigenGAN, i.e.,

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_i[\left\|\mathbf{U}_i^{\text{T}}\mathbf{U}_i - \mathbf{I}\right\|_1]. \tag{4}$$

Finally, output image is rendered by volume rendering technique [23]. At inference phase, we can control the discovered semantic attributes by manipulating corresponding element in $\mathbf{z}$. In addition, SURF-GAN enables explicit control over pose using viewing direction $\mathbf{v}$ as like other NeRF-based models.

3.2 Explicit control over pose with StyleGAN

In Sec. 3.2 and Sec. 3.3, we introduce a method to inject 3D perception and attribute controllability of SURF-GAN into StyleGAN.

Leveraging 3D-aware SURF-GAN. The first step is to transform pre-trained StyleGAN into a 3D controllable generator. We start with a question: How can we make StyleGAN be capable of controlling over pose explicitly when given arbitrary latent code? To this end, we utilize SURF-GAN as a pseudo ground-truth generator. It provides three images, i.e., I_s, I_c, I_t which denote source, canonical, and target image respectively. Here, $\mathbf{z}$ is fixed in all images but the view directions of I_s and I_t are randomly sampled and I_c has canonical view (i.e., $\mathbf{v}$=$[0, 0]$). Therefore, we can exploit them as multi-view supervision of the same identity. Afterwards, the images are embedded to $\mathcal{W}+$ [1] space by a GAN inversion encoder E, i.e, $\{\mathbf{w}_s, \mathbf{w}_c, \mathbf{w}_t\} = \{E(I_s), E(I_c), E(I_t)\}$. Here, we exploit the pre-trained pSp [47] encoder and it actually predicts the residual and adds it to the mean latent vector, but we omit the notation for simplicity.

Mapping to a canonical latent vector. To handle arbitrary pose without employing off-the-shelf 3D models, we need to build an additional process. To this end, we propose a canonical latent mapper T, which converts an arbitrary code to a canonical code in the latent space of StyleGAN. Here, the canonical code implies being a canonical pose (frontal) in image space. T takes $\mathbf{w}_s$ as input and predicts its frontalized version $\hat{\mathbf{w}}_c = T(\mathbf{w}_s)$ with the mapping function. In order to train T, we exploit latent loss to minimize the difference between the predicted $\hat{\mathbf{w}}_c$ and pseudo ground truth of canonical code $\mathbf{w}_c$, i.e.,

$$\mathcal{L}_{\mathrm{w}}^{c} = \|\mathbf{w}_c - T(\mathbf{w}_s)\|_1. \tag{5}$$

To guarantee plausible translation result in image space, we also adopt pixel-level ℓ_2-loss and LPIPS loss [63] between two decoded images, i.e.,

$$\mathcal{L}_{\mathrm{I}}^{c} = \|I'_c - \hat{I}_c\|_2^2 \tag{6}$$

$$\mathcal{L}_{\mathrm{LPIPS}}^{c} = \|F(I'_c) - F(\hat{I}_c)\|_2^2, \tag{7}$$

where I'_c and $\hat{I}_c$ represent the decoded images from $\mathbf{w}_c$ and $\hat{\mathbf{w}}_c$ respectively, and $F(\cdot)$ denotes the perceptual feature extractor. Hence, the loss for canonical view generation is formulated by

$$\mathcal{L}^{\mathrm{c}} = \lambda_1 \mathcal{L}_{\mathrm{w}}^{\mathrm{c}} + \lambda_2 \mathcal{L}_{\mathrm{I}}^{\mathrm{c}} + \lambda_3 \mathcal{L}_{\mathrm{LPIPS}}^{\mathrm{c}}. \tag{8}$$

Target view generation. Next, the canonical vector is converted to a target latent vector according to given a target view $\mathbf{v}_t = [\alpha, \beta]$ as an additional input. Here, α and β stand for pitch and yaw respectively. The manipulation is conducted in the latent space of StyleGAN by adding a pose vector which is obtained by a linear combination of pitch and yaw vectors ($\mathbf{p}$ and $\mathbf{y}$, respectively)

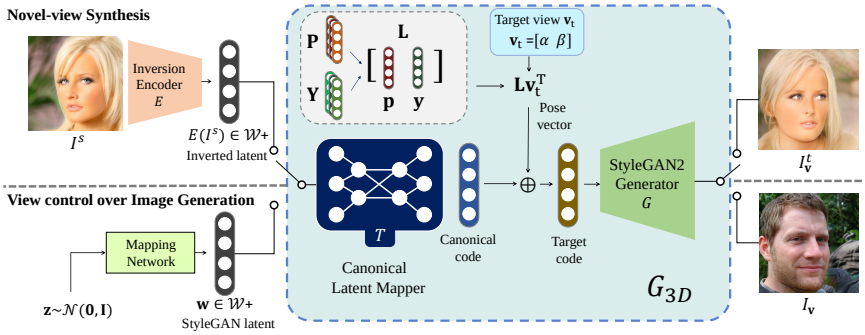

Fig. 4: The controllable StyleGAN allows explicit control over camera view. It can be used for novel pose synthesis (upper) and view-conditioned image generation (lower).

with $\mathbf{v}_t$ as coefficients, i.e., $\hat{\mathbf{w}}_t = \hat{\mathbf{w}}_c + \mathbf{L}\mathbf{v}_t^T$, where $\mathbf{L} = [\mathbf{p}\ \mathbf{y}]$. Therefore, we need to find optimal solution of $\mathbf{L}$ which can represent an adequate 3D control over pose. Although earlier studies [50,40] have shown successful interpolation results with the linear manipulation, unfortunately, they have found sub-optimal solutions that just control the intended pose attribute implicitly rather than explicit control over 3D camera. The interesting fact we observed is that the pose-related attribute (e.g., yaw) is not uniquely determined by a single direction. Rather, several orthogonal directions can have different effects on the same attribute. For example, two orthogonal direction A and B both can affect yaw but work differently. Based on this observation, we exploit several sub-direction vectors to compensate marginal portion that is not captured by a single direction vector. Our hypothesis is that the optimal direction that follows real geometry can be obtained by a proper combination of the sub-direction vectors. Borrowing the idea of basis in Sec. 3.1, we construct each of N learnable basis to obtain final pose vectors for pitch and yaw respectively. Therefore, we optimize the matrices $\mathbf{P} = [\mathbf{d}_1^p, \ldots, \mathbf{d}_N^p]$ and $\mathbf{Y} = [\mathbf{d}_1^y, \ldots, \mathbf{d}_N^y]$. The process to obtain the target vector can be described as,

$$\hat{\mathbf{w}}_t = \hat{\mathbf{w}}_c + \sum_{i=1}^{N} (\alpha \cdot l_i^p \mathbf{d}_i^p + \beta \cdot l_i^y \mathbf{d}_i^y), \tag{9}$$

where the l_i^p and l_i^y represent the learnable scalining factor deciding the importance of basis $\mathbf{d}_i^p$ and $\mathbf{d}_i^y$ respectively. To penalize finding redundant directions, we add orthogonal regularization, i.e.,

$$\mathcal{L}_{\text{reg}} = \left\|\mathbf{P}^{\mathrm{T}}\mathbf{P} - \mathbf{I}\right\|_1 + \left\|\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{I}\right\|_1. \tag{10}$$

Similar to the canonical view generation, the model is penalized by the difference of the latent codes ($\mathbf{w}_t$ vs. $\hat{\mathbf{w}}_t$) and that of the corresponding decoded images (I_t' vs. $\hat{I}_t$). In addition, we also utilize LPIPS loss. Therefore, the objective function of target view generation is described as,

$$\mathcal{L}^{\mathrm{t}} = \lambda_4 \mathcal{L}_{\mathrm{w}}^{\mathrm{t}} + \lambda_5 \mathcal{L}_{\mathrm{I}}^{\mathrm{t}} + \lambda_6 \mathcal{L}_{\mathrm{LPIPS}}^{\mathrm{t}} + \lambda_7 \mathcal{L}_{\mathrm{reg}}. \tag{11}$$

Finally, the full objective to train the proposed modules can be formulated as $\mathcal{L} = \mathcal{L}^{\mathrm{c}} + \mathcal{L}^{\mathrm{t}}$. After training, StyleGAN (G) becomes a 3D-controllable generator (G_{3D}) with the proposed modules as illustrated in Fig. 4. We can achieve a high quality image with intended pose by conditioning view as follow,

$$I_{\mathbf{v}} = G_{3D}(\mathbf{w}, \mathbf{v}_t) = G(\mathbf{w} + T(\mathbf{w}) + \mathbf{L}\mathbf{v}_t^T), \tag{12}$$

where $I_{\mathbf{v}}$ represents a generated image with target pose $\mathbf{v}_t$ and $\mathbf{w} \in \mathcal{W}+$ is duplicated version of 512-dimensional style vector in $\mathcal{W}$ which is obtained by the mapping network in StyleGAN. Moreover, we can extend our method to synthesize novel view of real images by combining with GAN inversion, i.e.,

$$I_{\mathbf{v}}^t = G_{3D}(E(I^s), \mathbf{v}_t), \tag{13}$$

where I^s is an input source image in arbitrary view and $I_{\mathbf{v}}^t$ denotes a generated target image with target pose $\mathbf{v}_t$. Note that our method can handle arbitrary images without exploiting off-the-shelf 3D models such as pose detectors or 3D fitting models. In addition, it synthesizes output at once without an iterative optimization process for overfitting latent code into an input portrait image.

3.3 Finding semantic direction with SURF-GAN

Beyond 3D perception, we can discover semantic directions in the latent space of StyleGAN that can control facial attributes using SURF-GAN generated images. Such directions can be obtained by a simple vector arithmetic [46] with two latent codes or several interpolated samples generated by SURF-GAN. Although our approach does not overwhelm state-of-the-art methods analyzing via supervision, it would be a simple yet effective alternative that can provide pose-robust editing directions. Of course, the discovery using SURF-GAN is one of many applicable approaches and we can also utilize the existing semantic analysis methods [50,17,51] because our model is flexibly compatible with well-studied StyleGAN-based techniques.

4 Experimental result

This section presents qualitative and quantitative comparisons with state-of-the-art methods and analysis of our method. Additional experiments and discussions not included in this paper can be found in the supplementary material.

4.1 Implementation

SURF-GAN. We use each of two datasets to train SURF-GAN, i.e., CelebA [34] dataset and FFHQ [28] dataset. We set the number of sub-modulations in each layer $K = 6$ (Eq. 2 and Eq. 3) and the number of modulated layers (SURF blocks) is nine ($\therefore t = 8$). The other settings are roughly the same with those of π-GAN. More details can be found in the supplementary paper.

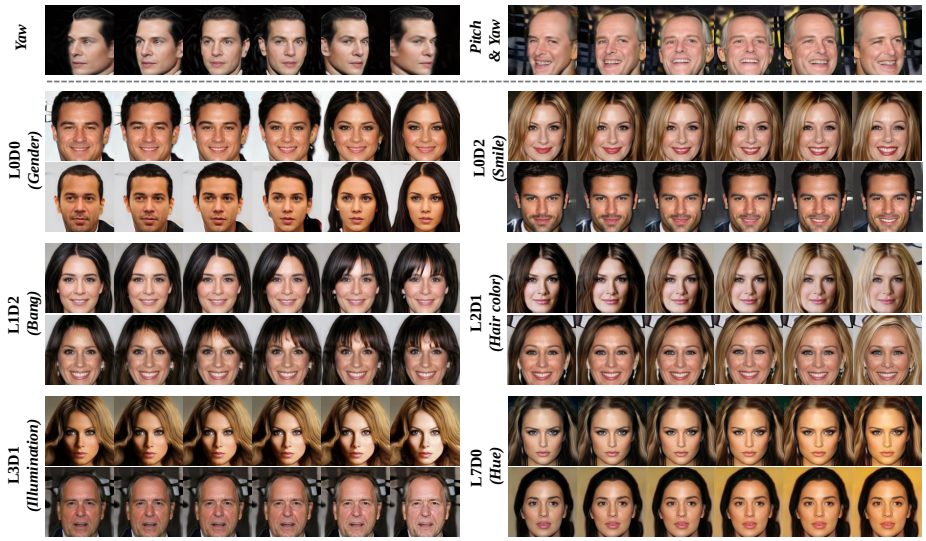

Fig. 5: Discovered semantic attributes at different layers in SURF-GAN. We can manipulate the attributes (e.g., hair color, gender, etc.) with the control parameters as well as explicit control over 3D camera. LiDj denotes the j^{th} basis of the layer i^{th}.

3D-controllable StyleGAN. For training of 3D controllable StyleGAN, we exploit generated images by SURF-GAN trained with FFHQ because StyleGAN [29] and GAN inversion encoder [47] are pre-trained with FFHQ. We design the model to alter only the first four $\mathbf{w}$ vectors (i.e., 4×512) which have been known to control pose [28,64]. We set the number of sub-direction $N = 5$ (Eq. 9). The hyper-parameter of the loss function (Eq. 8 and Eq. 11) are set to λ_1,λ_4=10, $\lambda_7 = 100$, and 1.0 for the others.

4.2 Controllability of SURF-GAN

First, we present the attributes of CelebA discovered by SURF-GAN in Fig. 5. As like other 3D-aware GANs [8,39,49,66,16,7], it can synthesis a view-conditioned image, i.e., yaw and pitch can be controlled explicitly with input view direction (top row). In contrast to other 3D NeRF-GANs, SURF-GAN can discover semantic attributes in different layers in an unsupervised manner. Additionally, the discovered attributes can be manipulated by the corresponding control parameters. As shown in Fig. 5, different layers of SURF-GAN capture diverse attributes such as gender, hair color, illumination, etc. Interestingly, we observe the early layers capture high-level semantics (e.g., overall shape or gender) and the rear layers focus fine details or texture (e.g., illumination or hue). This property is similar to that seen in 2D GANs even though SURF-GAN consists of MLPs without convolutional layers. Additional discovered attributes, those of FFHQ and the comparison with π-GAN, which is a pure NeRF-GAN as like ours can be found in the supplementary material.

Table 1: Quantitative comparison of the proposed 3D controllable StyleGAN with other 3D controllable generative models. We use FID, pose accuracy, and frames per second for evaluation. † denotes quoting from the original paper.

	ConfigNet	π-GAN	CIPS-3D	LiftedGAN	Ours
FID (↓)	33.41†	47.68	6.97†	29.81†	**4.72**
Pose err.($\times 10^{-2}$) (↓)	9.56	**3.81**	9.12	5.52	4.24
Frames/Sec. (↑)	**345**	4	22	56	72

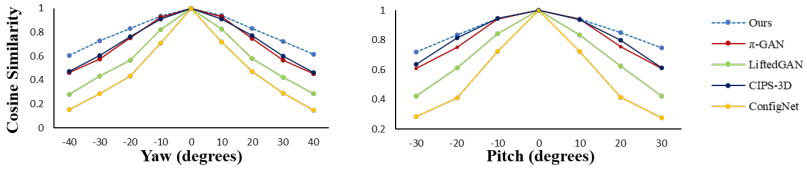

Fig. 6: Quantitative comparison of 3D-controllable models on identity preservation under different angles using the averaged cosine similarity from ArcFace [11].

4.3 Portrait image generation with 3D control

To evaluate the performance of the proposed 3D-controllable StyleGAN, we report the qualitative and quantitative comparison with state-of-the-art models [30,8,66,52] whose generator allows explicit control over pose. Fig. 7 shows synthesis results of each model for given target views. Here, the results are 256^2 images generated by each method trained with FFHQ [28]. ConfigNet reveals lack of visual quality and weakness in large pose changes. π-GAN shows the accurate geometry because its generator consists of pure NeRF layers, but this property also results in some degenerated visual quality. CIPS-3D presents improved visual quality by adopting followed 2D INR network, but it suffers from 3D inaccuracy in specific poses. LiftedGAN generates reasonable outputs according to target views by utilizing differentiable renderer, but it lacks photorealism. Our method generates photorealistic images and shows plausible control over pose and multi-view consistency. We also report the quantitative comparisons of the models in Table. 1 and Fig. 6. We use FID score, pose accuracy estimated by 3D model [69], frames per second, and identity similarity [11] as evaluation metrics. Compared to 3D-aware models, our method achieves a competitive score on pose accuracy and delivers superior results in efficiency, visual quality, and multi-view consistency. Although 2D-based ConfigNet shows overwhelming efficiency, it struggles with multi-view consistency and photorealism.

4.4 Novel view synthesis of real image

By utilizing GAN inversion method, our method can perform novel view synthesis from a single portrait. Here, we use pSp [47] encoder for the inversion.

Fig. 7: Qualitative results of 3D-controllable generative models under target poses.

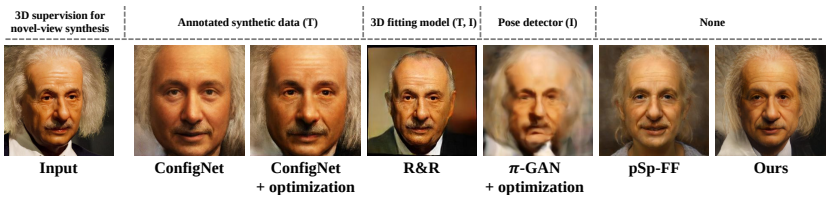

Fig. 8: Face frontalization results by the methods that can edit the pose of real image. Upper row denotes 3D supervision of each method for training (T) and inference (I).

To demonstrate the effectiveness of the canonical mapper, we firstly present the frontalization results in Fig. 8, which is a special case of novel pose synthesis. We mark the 3D supervision in training and inference for each method. Here, pSp-FF denotes the frontalization-only version of pSp [47]. Our method successfully generates a canonical view while preserving the identity. Next, we further compare the novel view synthesis results of each model. ConfigNet and π-GAN with optimizing latent code through iterative manner for overfitting to single test image show inferior results, especially in large pose variation. Rotate-and-Render (R&R) [65] presents reasonable results by exploiting off-the-shelf 3D fitting models [69] in the generation process. However, R&R loses some fine details of properties of the original, such as hair or background. Our models can edit pose successfully while preserving identity even though it does not require off-the-shelf 3D models and additional optimization for overfitting to an input. It is also demonstrated by the quantitative results in Fig. 9 which reports the averaged cosine similarity between input image and outputs at given various

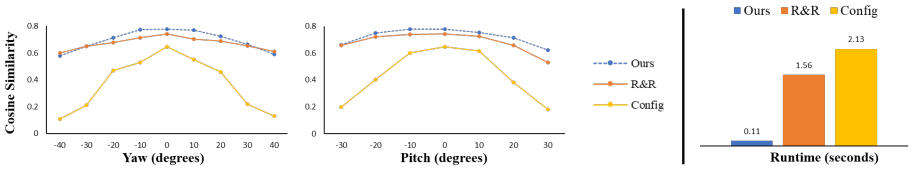

Fig. 9: Quantitative results of novel view synthesis models. We compute identity similarity between input and synthesized images using ArcFace (left) and runtime (right).

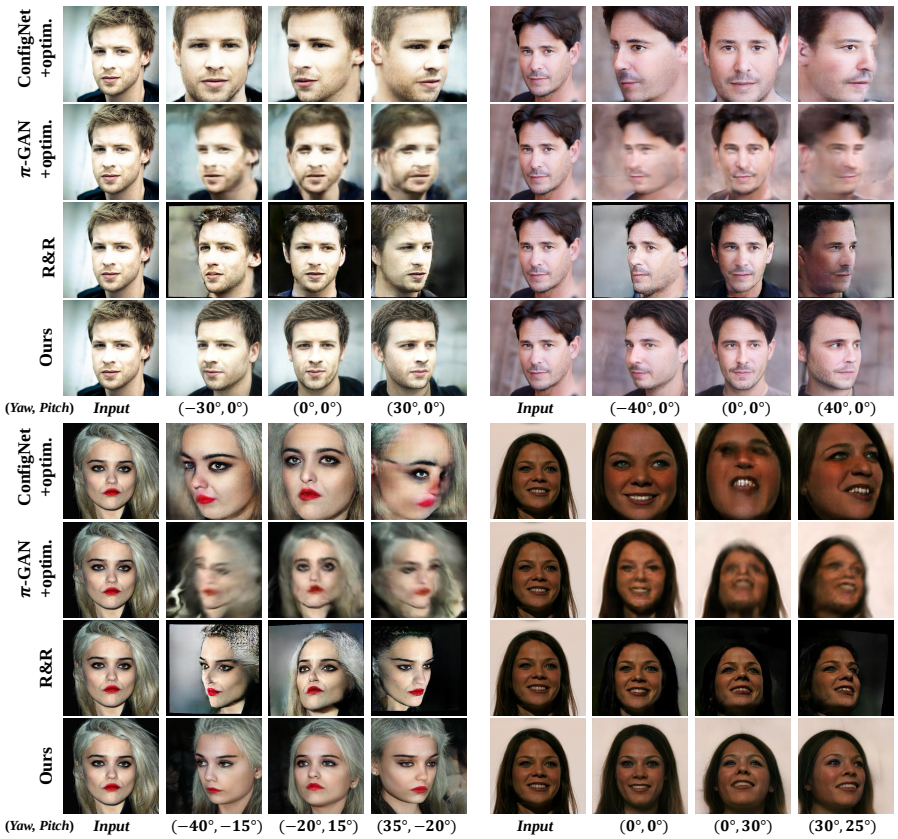

Fig. 10: Results of novel view synthesis under various target poses using CelebA-HQ.

angles using ArcFace [11] and runtime of each method to process a single image.

4.5 Semantic attribute manipulation under conditioned poses

Fig. 11 presents the results of semantic attribute editing with pose control by 3D-controllable StyleGAN. The upper row stands for controllable generation (a)

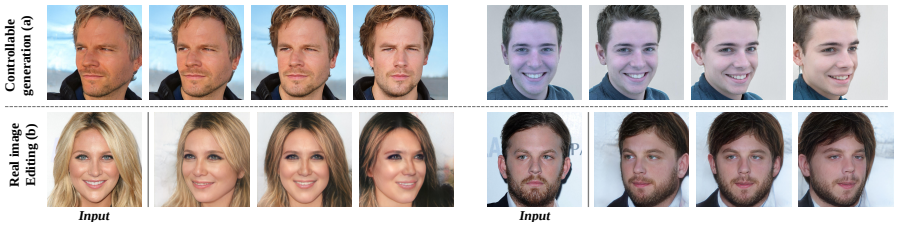

Fig. 11: Editing both attributes and view direction by 3D-controllable StyleGAN.

Fig. 12: Explicit 3D control over real and stylized images.

and the lower represents real image editing (b). The presented attributes, i.e., skin color, hue, hair color, and bangs are those discovered by SURF-GAN.

4.6 Applications

Our model can be flexibly integrated with other methods that also exploit pre-trained StyleGAN. Beyond the real image domain, we present a novel view synthesis of the stylized images such as toon or painting in Fig. 12. We use a interpolated StyleGAN proposed by Pinkney and Alder [45] for toonifying and a transferred StyleGAN trained with MetFace [26] for painting-style outputs.

5 Conclusion

In this paper, we solved the problems of 3D-aware GANs and 2D GANs by introducing SURF-GAN and 3D-controllable StyleGAN. Unlike other 3D-aware GANs, SURF-GAN can discover meaningful semantics and control them in an unsupervised manner. Using SURF-GAN, we convert StyleGAN to be explicitly 3D-controllable and it delivers outstanding results in both random image generation and novel view synthesis of real image. In addition, our method has the potential to be flexibly combined with other methods. We expect our work will be used practically and effectively in various tasks and hope it will open up a new direction in 3D-aware generation and editing fields.

Acknowledgement. This work was supported by DMLab. We also thank to Anonymous ECCV Reviewers for their constructive suggestions and discussions on our paper.

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 4, 7
2. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 4
3. Abdal, R., Zhu, P., Mitra, N.J., Wonka, P.: Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. ACM Transactions on Graphics (TOG) (2021) 4
4. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. In: International Conference on Computer Vision (ICCV) (2021) 4
5. Alaluf, Y., Tov, O., Mokady, R., Gal, R., Bermano, A.H.: Hyperstyle: Stylegan inversion with hypernetworks for real image editing. arXiv (2021) 4
6. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (ICLR) (2019) 1, 4
7. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 2, 4, 5, 10
8. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2, 3, 4, 5, 10, 11
9. Chen, A., Liu, R., Xie, L., Chen, Z., Su, H., Yu, J.: Sofgan: A portrait image generator with dynamic styling. Transactions on Graphics (TOG) (2022) 4
10. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS) (2016) 4
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 11, 13
12. Deng, Y., Yang, J., Chen, D., Wen, F., Tong, X.: Disentangled and controllable face image generation via 3d imitative-contrastive learning. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 2, 4
13. Deng, Y., Yang, J., Xiang, J., Tong, X.: GRAM: Generative radiance manifolds for 3d-aware image generation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 2, 4
14. Gadelha, M., Maji, S., Wang, R.: 3d shape induction from 2d views of multiple objects. In: International Conference on 3D Vision (3DV) (2017) 4
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS) (2014) 1
16. Gu, J., Liu, L., Wang, P., Theobalt, C.: StyleNeRF: A style-based 3d-aware generator for high-resolution image synthesis. arXiv (2021) 2, 4, 5, 10

17. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. arXiv (2020) 3, 4, 9
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 6
19. He, Z., Kan, M., Shan, S.: EigenGAN: Layer-wise eigen-learning for gans. In: International Conference on Computer Vision (ICCV) (2021) 4, 5
20. Henzler, P., Mitra, N.J., Ritschel, T.: Escaping plato's cave: 3d shape from adversarial rendering. In: International Conference on Computer Vision (ICCV) (2019) 4
21. Hu, Y., Wu, X., Yu, B., He, R., Sun, Z.: Pose-guided photorealistic face rotation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4
22. Jeon, I., Lee, W., Pyeon, M., Kim, G.: Ib-gan: Disengangled representation learning with information bottleneck generative adversarial networks. In: AAAI Conference on Artificial Intelligence (AAAI) (2021) 4
23. Kajiya, J.T., Von Herzen, B.P.: Ray tracing volume densities. SIGGRAPH (1984) 5, 6
24. Kaneko, T., Hiramatsu, K., Kashino, K.: Generative attribute controller with conditional filtered generative adversarial networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 4
25. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (ICLR) (2018) 1, 4
26. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) 14
27. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2021) 1, 4
28. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1, 4, 5, 9, 10, 11
29. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 1, 4, 5, 10
30. Kowalski, M., Garbin, S.J., Estellers, V., Baltrušaitis, T., Johnson, M., Shotton, J.: Config: Controllable neural face image generation. In: European Conference on Computer Vision (ECCV) (2020) 2, 4, 11
31. Kwak, J.g., Li, Y., Yoon, D., Han, D., Ko, H.: Generate and edit your own character in a canonical view. arXiv (2022) 4
32. Lee, W., Kim, D., Hong, S., Lee, H.: High-fidelity synthesis with disentangled representation. In: European Conference on Computer Vision (ECCV) (2020) 4
33. Liao, Y., Schwarz, K., Mescheder, L., Geiger, A.: Towards unsupervised learning of generative models for 3d controllable image synthesis. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 4
34. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 9
35. Lunz, S., Li, Y., Fitzgibbon, A., Kushman, N.: Inverse graphics gan: Learning to generate 3d shapes from unstructured 2d data. arXiv (2020) 4
36. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision (ECCV) (2020) 2, 5

37. Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., Yang, Y.L.: HoloGAN: Unsupervised learning of 3d representations from natural images. In: International Conference on Computer Vision (ICCV) (2019) 2, 4
38. Nguyen-Phuoc, T., Richardt, C., Mai, L., Yang, Y.L., Mitra, N.: BlockGAN: Learning 3d object-aware scene representations from unlabelled images. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) 2, 4
39. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2, 4, 5, 10
40. Nitzan, Y., Gal, R., Brenner, O., Cohen-Or, D.: Large: Latent-based regression through gan semantics. arXiv (2021) 4, 8
41. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: StyleSDF: High-resolution 3d-consistent image and geometry generation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 4
42. Pan, X., Dai, B., Liu, Z., Loy, C.C., Luo, P.: Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. In: International Conference on Learning Representations (ICLR) (2021) 2, 4
43. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: International Conference on Computer Vision (ICCV) (2021) 4
44. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: Film: Visual reasoning with a general conditioning layer. In: AAAI Conference on Artificial Intelligence (AAAI) (2018) 6
45. Pinkney, J.N., Adler, D.: Resolution dependent gan interpolation for controllable image synthesis between domains. arXiv (2020) 14
46. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv (2015) 9
47. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4, 7, 10, 11, 12
48. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. arXiv (2021) 4, 5
49. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: GRAF: Generative radiance fields for 3d-aware image synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) 2, 4, 10
50. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 3, 4, 8, 9
51. Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in gans. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 3, 4, 9
52. Shi, Y., Aggarwal, D., Jain, A.K.: Lifting 2d stylegan for 3d-aware face generation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2, 4, 11
53. Shoshan, A., Bhonker, N., Kviatkovsky, I., Medioni, G.: Gan-control: Explicitly controllable gans. In: International Conference on Computer Vision (ICCV) (2021) 4
54. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) 6

55. Sun, J., Wang, X., Zhang, Y., Li, X., Zhang, Q., Liu, Y., Wang, J.: Fenerf: Face editing in neural radiance fields. arXiv (2021) 4
56. Szabó, A., Meishvili, G., Favaro, P.: Unsupervised generative 3d shape learning from natural images. arXiv (2019) 4
57. Tewari, A., Elgharib, M., Bernard, F., Seidel, H.P., Pérez, P., Zollhöfer, M., Theobalt, C.: Pie: Portrait image embedding for semantic control. ACM Transactions on Graphics (TOG) (2020) 2, 4
58. Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H.P., Perez, P., Zollhofer, M., Theobalt, C.: Stylerig: Rigging stylegan for 3d control over portrait images. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 2, 4
59. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) (2021) 4
60. Xue, Y., Li, Y., Singh, K.K., Lee, Y.J.: GIRAFFE HD: A high-resolution 3d-aware generative model. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 2, 4
61. Yao, X., Newson, A., Gousseau, Y., Hellier, P.: A latent transformer for disentangled face editing in images and videos. In: International Conference on Computer Vision (ICCV) (2021) 4
62. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Towards large-pose face frontalization in the wild. In: International Conference on Computer Vision (ICCV) (2017) 4
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 7
64. Zhang, Y., Chen, W., Ling, H., Gao, J., Zhang, Y., Torralba, A., Fidler, S.: Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. arXiv (2020) 10
65. Zhou, H., Liu, J., Liu, Z., Liu, Y., Wang, X.: Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 12
66. Zhou, P., Xie, L., Ni, B., Tian, Q.: Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. arXiv (2021) 2, 4, 5, 10, 11
67. Zhu, J., Shen, Y., Zhao, D., Zhou, B.: In-domain gan inversion for real image editing. In: European Conference on Computer Vision (ECCV) (2020) 4
68. Zhu, J.Y., Zhang, Z., Zhang, C., Wu, J., Torralba, A., Tenenbaum, J.B., Freeman, W.T.: Visual Object Networks: Image generation with disentangled 3D representations. In: Advances in Neural Information Processing Systems (NeurIPS) (2018) 4
69. Zhu, X., Liu, X., Lei, Z., Li, S.Z.: Face alignment in full pose range: A 3d total solution. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2017) 11, 12