Self-Supervised Geometry-Aware Encoder for Style-Based 3D GAN Inversion

Yushi Lan[1] Xuyi Meng[1] Shuai Yang[1] Chen Change Loy[1] ✉ Bo Dai[2]
[1]S-Lab, Nanyang Technological University, Singapore [2]Shanghai AI Laboratory

Abstract

StyleGAN has achieved great progress in 2D face reconstruction and semantic editing via image inversion and latent editing. While studies over extending 2D StyleGAN to 3D faces have emerged, a corresponding generic 3D GAN inversion framework is still missing, limiting the applications of 3D face reconstruction and semantic editing. In this paper, we study the challenging problem of 3D GAN inversion where a latent code is predicted given a single face image to faithfully recover its 3D shapes and detailed textures. The problem is ill-posed: innumerable compositions of shape and texture could be rendered to the current image. Furthermore, with the limited capacity of a global latent code, 2D inversion methods cannot preserve faithful shape and texture at the same time when applied to 3D models. To solve this problem, we devise an effective self-training scheme to constrain the learning of inversion. The learning is done efficiently without any real-world 2D-3D training pairs but proxy samples generated from a 3D GAN. In addition, apart from a global latent code that captures the coarse shape and texture information, we augment the generation network with a local branch, where pixel-aligned features are added to faithfully reconstruct face details. We further consider a new pipeline to perform 3D view-consistent editing. Extensive experiments show that our method outperforms state-of-the-art inversion methods in both shape and texture reconstruction quality.

1. Introduction

This work aims to devise an effective approach for encoder-based 3D Generative Adversarial Network (GAN) inversion. In particular, we focus on the reconstruction of 3D face, requiring just a single 2D face image as the input. In the inversion process, we wish to map a given image to the latent space and obtain an editable latent code with an encoder. The latent code will be further fed to a generator to reconstruct the corresponding 3D shape with high-quality shape and texture. Further to the learning of an inversion encoder, we also wish to develop an approach to synthesize 3D view-consistent editing results, *e.g.*, changing a neutral expression to smiling, by altering the estimated latent code.

GAN inversion [50] has been extensively studied for 2D images but remains underexplored in the 3D world. Inversion can be achieved via optimization [1, 2, 41], which typically provides a precise image-to-latent mapping but can be time-consuming, or encoder-based techniques [40, 47, 49], which explicitly learn an encoding network that maps an image into the latent space. Encoder-based techniques enjoy faster inversion, but the mapping is typically inferior to optimization. In this study, we extend the notion of encoder-based inversion from 2D images to 3D shapes.

Adding the additional dimension makes inversion more challenging beyond the goal of reconstructing an editable shape with detail preservation. In particular, **1)** Recovering 3D shapes from 2D images is an ill-posed problem, where innumerable compositions of shape and texture could generate identical rendering results. 3D supervisions are crucial to alleviate the ambiguity of shape inversion from images. Though high-quality 2D datasets are easily accessible, owing to the expensive cost of scans there is currently a lack of large-scale labeled 3D datasets. **2)** The global latent code, due to its compact and low-dimensional nature, only captures the coarse shape and texture information. Without high-frequency spatial details, we cannot generate high-fidelity outputs. **3)** Compared with 2D inversion methods where the editing view mostly aligns with the *input view*, in 3D editing we expect the editing results to perform well over the *novel views* with large pose variations. Therefore, 3D GAN inversion is non-trivial task and could not be achieved by directly applying existing approaches.

To this end, we propose a novel **E**ncoder-based **3D G**AN inv**E**rsion framework, E3DGE, which addresses the aforementioned three challenges. Our framework has three novel components with a delicate model design. Specifically:

Learning Inversion with Self-supervised Learning - The first component focuses on the training of the inversion encoder. To address the shape collapse of single-view 3D reconstruction without external 3D datasets, we retrofit the generator of a 3D GAN model to provide us with diverse pseudo training samples, which can then be used to train our inversion encoder in a self-supervised manner. Specifically, we generate 3D shapes from the latent space $\mathcal{W}$ of a 3D GAN, and then render diverse 2D views from each 3D shape given different camera poses. In this way, we can generate many pseudo 2D-3D pairs together with the corresponding latent codes. Since the pseudo pairs are generated from a smooth latent space that learns to approximate a nat-

ural shape manifold, they serve as effective surrogate data to train the encoder, avoiding potential shape collapse.

Local Features for High-Fidelity Inversion - The second component learns to reconstruct accurate texture details. Our novelty here is to leverage local features to enhance the representation capacity, beyond just the global latent code generated by the inversion encoder. Specifically, in addition to inferring an editable global latent code to represent the overall shape of the face, we further devise an hour-glass model to extract local features over the residuals details that the global latent code fails to capture. The local features, with proper projection to the 3D space, serve as conditions to modulate the 2D image rendering. Through this effective learning scheme, we marry the benefits of both global and local priors and achieve high-fidelity reconstruction.

Synthesizing View-consistent Edited Output - The third component addresses the problem of novel view synthesis, a problem unique to 3D shape editing. Specifically, though we achieve high-fidelity reconstruction through aforementioned designs, the local residual features may not fully align with the scene when being semantically edited. Moreover, the occlusion issue further degrades the fusion performance when rendering from novel views with large pose variations. To this end, we propose a 2D-3D hybrid alignment module for high-quality editing. Specifically, a 2D alignment module and a 3D projection scheme are introduced to jointly align the local features with edited images and inpaint occluded local features in novel view synthesis.

Extensive experiments show that our method achieves 3D GAN inversion with plausible shapes and high-fidelity image reconstruction without affecting editability. Owing to the self-supervised training strategy with delicate global-local design, our approach performs well on real-world 2D and 3D benchmarks without resorting to any real-world 3D dataset for training. To summarize, our main contributions are as follows:

- We propose an early attempt at learning an encoder-based 3D GAN inversion framework for high-quality shape and texture inversion. We show that, with careful design, samples synthesized by a GAN could serve as proxy data for self-supervised training in inversion.
- We present an effective framework that uses local features to complement the global latent code for high-fidelity inversion.
- We propose an effective approach to synthesize view-consistent output with a 2D-3D hybrid alignment.

2. Related Work

3D-aware Image Synthesis. Generative Adversarial Network [13] has shown promising results in generating photorealistic images [5, 21, 22] and inspired researchers to put efforts on 3D aware generation [15, 31, 34]. However, these methods use explicit shape representations, *i.e.*, voxels [15, 31] and meshes [34] as the intermediate shape models, which lacks photorealism and is memory-inefficient. Motivated by the recent success of neural rendering [28, 29, 37], researchers shift to implicit function along with the volume rendering process as the incorporated 3D inductive bias. Especially, NeRF [29] proposed an implicit 3D representation for novel view synthesis which defines a scene as $\{\boldsymbol{c}, \sigma\} = F_\Phi(\boldsymbol{x}, \boldsymbol{v})$, where $\boldsymbol{x}$ is the query point, $\boldsymbol{v}$ is the viewing direction from camera origin to $\boldsymbol{x}$, $\boldsymbol{c}$ is the emitted radiance (color) and σ is the volume density. Researchers further extend NeRF to generation task [7, 45] and show impressive 3D-awareness synthesis. To increase the generation resolution, recent works [8, 16, 52] resort to voxel-based representations or adopting a hybrid design [8, 14, 32, 33]. By lifting the intermediate low-resolution 2D features to high resolution with a 2D super-resolution decoder, the hybrid design achieves high resolution of 1024^2.

GAN-supervised Training. Previous works [4, 17, 18, 27, 36, 53, 55] propose to use pretrained GAN to generate training dataset. Through careful design in the sampling strategy [18], loss functions [36] and generation process [55], researches show that off-the-shelf image generators could facilitate a series of downstream visual applications.

2D GAN Inversion. Optimization-based 2D GAN inversion methods [1, 12] achieve photorealistic reconstruction at the cost of slow inference and lack of editability. To speed up, Encoder-based methods [9, 40, 47, 49, 56] like pSp [40] and e4e [47] have been developed and show better properties in editing through specific model design [40, 49] and training strategies [47]. However, they [1, 2, 40, 47, 56] all adopt global latent code alone for GAN inversion task, thus failing to recover high-fidelity details. Recently, HFGI [49] introduce an extra spatial consultation map to mitigate this issue, though still designed to restore 2D textures without considering 3D shape modeling. In this work, we propose a delicate design that exploits local features to recover texture details and achieves view-consistent synthesis.

3D GAN Inversion and Editing. Recent development of 3D GANs [7, 8, 14, 32, 33, 45] also calls for corresponding inversion frameworks. π-GAN and EG3D [8] directly adopt 2D inversion method [1, 41], which requires expensive latent or model optimization and still introduces implausible shape artifacts. The most relevant work to ours is Lin *et al.* [26], which employs a computationally expensive optimization-based framework [1] and combines FLAME [11, 25] for portrait animation. However, it fails to guarantee reasonable shape and is limited to human face domain. In parallel, some works are not based on 3D GAN and introduce a feed-forward [6, 11] or auto-decoder [39] pipeline for single-view 3D reconstruction [6, 11] or editing [39], which cannot leverage the strong GAN priors for high-resolution and flexible latent-based editing.

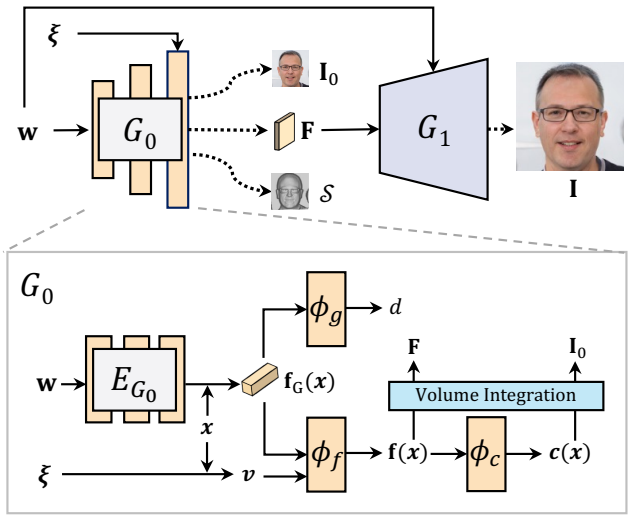


Figure 1. **StyleSDF.** Given a sampled latent code $\mathbf{w}$ and a camera pose $\boldsymbol{\xi}$, StyleSDF generates object SDF d to depict the shape and the corresponding face image $\mathbf{I}$.

3. Preliminaries

Hybrid 3D-aware Generation. To achieve high-resolution novel view synthesis, hybrid 3D-aware generator [8, 14, 32, 33] is proposed. It is a cascade model $G = G_1 \circ G_0$ composed of a NeRF-based renderer G_0 [7] and a 2D super-resolution network G_1, as shown in Fig. 1. Both G_0 and G_1 follow the style-based architecture [21, 23] to accept a latent code $\mathbf{w}$ to control the style of the generated object. During generation, G_0 captures the underlying geometry with the full control of $\mathbf{w}$ and camera pose $\boldsymbol{\xi}$, and renders a low resolution image $\mathbf{I_0}$ and an intermediate feature map $\mathbf{F}$. Then, G_1 further upsamples $\mathbf{F}$ to obtain a high-resolution image $\mathbf{I}$ with added high-frequency details.

Among them, StyleSDF [33] introduces signed distance function (SDF) to serve as a proxy for the density function $\sigma(\boldsymbol{x})$ used for the volume rendering in NeRF. Specifically, StyleSDF uses G_0 to predict the distance $d(\boldsymbol{x}) = G_0(\mathbf{w}, \boldsymbol{x})$ between the query point $\boldsymbol{x}$ and the shape surface, where the density function $\sigma(\boldsymbol{x})$ can be transformed from $d(\boldsymbol{x})$ for NeRF [29] to render. The incorporation of SDF leads to higher-quality geometry in terms of expressiveness view-consistency and clear definition of the surface. StyleSDF also enjoys the flexible style control for semantic editing as in StyleGAN [21]. Therefore, in this paper we mainly use StyleSDF as the base model for GAN inversion study. Note that our method is not limited to StyleSDF and could be easily extended to other style-based 3D GAN variations.

4. E3DGE

An effective 3D GAN inversion shall be capable of **1)** reconstructing plausible 3D shape given single-view input, **2)** maintaining high-fidelity texture, and **3)** allowing view-consistent semantic edits. To achieve these goals, we propose the E3DGE framework with three novel components: In Sec. 4.1, we leverage 3D GAN to generate pseudo 2D-3D paired samples for 3D supervisions, and train an inversion encoder E_0 to estimate the latent of plausible 3D shapes from a 2D image; In Sec. 4.2, we train a local encoder E_1 to extract pixel-aligned features to enrich texture details for high-fidelity inversion; Finally, Sec. 4.3 introduces a hybrid alignment module for view-consistent semantic editing.

4.1. Self-supervised Inversion Learning

In this section, we propose to mitigate the lack of large-scale high-quality 2D-3D paired datasets by retrofitting pre-trained 3D GANs to provide pseudo samples for training our inversion encoder. We demonstrate the model trained from pseudo samples can rival and even outperform the methods learned from real data on the 3D GAN inversion task. We detail the process as follows.

Global Encoder for 3D GAN Inversion. With the style-based G, we build our encoder E_0 based on pSp [40] for inversion. Given a target image $\mathbf{I}$, E_0 predicts its latent code $\hat{\mathbf{w}} = E_0(\mathbf{I})$. Given the corresponding camera pose $\boldsymbol{\xi}$, the reconstructed image is obtained by $\tilde{\mathbf{I}} = G(\hat{\mathbf{w}}, \boldsymbol{\xi})$ to approximate $\mathbf{I}$. In addition, we would like its 3D shape predicted by G_0 to be plausible enough.

Distill 3D GANs as 3D Supervisions. Different compositions of shape and texture could lead to identical 2D-rendered images. 3D supervision is needed to alleviate such shape-texture ambiguity. In the lack of large-scale high-quality 2D-3D paired samples, we formulate GAN Inversion as a *self-training* task, where samples synthesized from itself are leveraged to boost the reconstruction fidelity in both 2D and 3D domains.

As shown in Fig 1, we synthesize paired 3D shape information $\mathcal{S}$ and 2D image $\mathbf{I}$ from latent code $\mathbf{w}$ and camera pose $\boldsymbol{\xi}$ using G to train E_0. To extract the 3D shape information $\mathcal{S}$ of each synthetic shape, we first sample a point set $\mathcal{P} = \{\mathcal{P}_\mathcal{O}, \mathcal{P}_\mathcal{F}\}$ where $\mathcal{P}_\mathcal{O}$ and $\mathcal{P}_\mathcal{F}$ contain points sampled from the surface and around the surface, respectively. Then, we calculate the geometry descriptor d_i and $\boldsymbol{n}_i$ for each 3D point $\boldsymbol{x}_i \in \mathcal{P}$, and $\mathcal{S}$ is defined as the set of geometry descriptors of all 3D point in $\mathcal{P}$:

$$\begin{aligned} \mathcal{S} = \{\{d_i, \boldsymbol{n}_i\}_{i=1}^{|\mathcal{P}|} \mid \\ \boldsymbol{x}_i \in \mathcal{P}, d_i = G_0(\mathbf{w}, \boldsymbol{x}_i), \boldsymbol{n}_i = \nabla_{\boldsymbol{x}_i} d_i\}, \end{aligned} \tag{1}$$

where d_i is the distance from $\boldsymbol{x}_i$ to the shape surface and $\boldsymbol{n}_i$ is the surface normal defined by the gradient of the distance w.r.t. $\boldsymbol{x}_i$. Note our method is not limited to the SDF-based shape representation and can be easily extended to radiance-based methods [7, 8, 35]. Moreover, given different camera poses, we can generate a diverse 2D-3D dataset to help alleviate the shape-texture ambiguity, *i.e.*, for each shape $\mathcal{S}$,

various images $\mathbf{I} = G(\mathbf{w}, \boldsymbol{\xi})$ can be rendered by randomly sampling $\boldsymbol{\xi}$ from a predefined pose distribution $p_{\boldsymbol{\xi}}$. Finally, we define $\mathcal{X} = \{\mathcal{S}, \boldsymbol{\xi}, \mathbf{I}\}$ as a training sample for $E_{\mathbf{0}}$.

3D GAN-Supervised Training. As shown in Fig. 2 (a), given a training sample $\mathcal{X}$, the forward process is represented as:

$$\hat{\mathbf{w}} = E_{\mathbf{0}}(\mathbf{I}) \tag{2}$$

$$\{\tilde{\mathbf{I}}, \hat{\mathcal{S}}\} = G(\hat{\mathbf{w}}, \boldsymbol{\xi}, \mathcal{P}) \tag{3}$$

where $\hat{\mathbf{w}}$ is the estimated latent code and $\hat{\mathcal{S}} = \{\{\hat{d}_i, \hat{\boldsymbol{n}}_i\}_{i=1}^{|\mathcal{P}|} \mid \boldsymbol{x}_i \in \mathcal{P}\}$ is the estimated 3D shape information conditioned on $\hat{\mathbf{w}}$ and $\mathcal{P}$.

To achieve 3D supervision, we would like the estimated $\hat{\mathcal{S}}$ to approximate the ground truth $\mathcal{S}$. Specifically, for points over the surface, their distances and normal are both considered while for points around the surface, we only supervise their distance following [3, 37], leading to geometry loss:

$$\mathcal{L}_{geo}^{\mathcal{O}} = \mathbb{E}_{\mathcal{X}}\left[\frac{1}{|\mathcal{P}_{\mathcal{O}}|}\sum_{i=1}^{|\mathcal{P}_{\mathcal{O}}|}\lambda_{g1}|\hat{d}_i| + \lambda_{g2}\|\hat{\boldsymbol{n}}_i - \boldsymbol{n}_i\|_1\right] \tag{4}$$

$$\mathcal{L}_{geo}^{\mathcal{F}} = \mathbb{E}_{\mathcal{X}}\left[\frac{1}{|\mathcal{P}_{\mathcal{F}}|}\sum_{i=1}^{|\mathcal{P}_{\mathcal{F}}|}\lambda_{g3}|\hat{d}_i - d_i|\right] \tag{5}$$

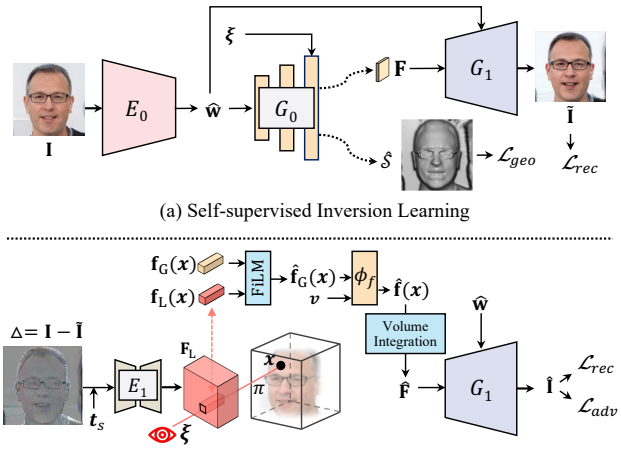
$$\mathcal{L}_{geo} = \mathcal{L}_{geo}^{\mathcal{O}} + \mathcal{L}_{geo}^{\mathcal{F}}, \tag{6}$$

where λs are loss weights and $d_i = 0$ for points over the surface. We also impose code reconstruction loss $\mathcal{L}_{code} = \|\hat{\mathbf{w}} - \mathbf{w}\|_2$ to regularize the learning and 2D supervisions $\mathcal{L}_{rec}$ to minimize the reconstruction error between $\tilde{\mathbf{I}}$ and $\mathbf{I}$ as in pSp [40]. The overall loss is $\mathcal{L} = \mathcal{L}_{geo} + \mathcal{L}_{code} + \mathcal{L}_{rec}$.

4.2. Local Features for High-Fidelity Inversion

To facilitate introductions in the following sections, we first take a look at the details of StyleSDF. As shown in Fig. 1, $G_{\mathbf{0}}$ can be further divided into four parts: a 8-layer MLP encoder $E_{G_{\mathbf{0}}}$, a SDF decoder ϕ_g, a feature decoder ϕ_f and a color decoder ϕ_c. $E_{G_{\mathbf{0}}}$ extracts a global feature $\mathbf{f}_{\text{G}}(\boldsymbol{x}) = E_{G_{\mathbf{0}}}(\boldsymbol{x}, \mathbf{w})$. Based on $\mathbf{f}_{\text{G}}$, ϕ_g and ϕ_f compute SDF $d(\boldsymbol{x}) = \phi_g(\mathbf{f}_{\text{G}}(\boldsymbol{x}))$ and the last-layer feature $\mathbf{f}(\boldsymbol{x}, \boldsymbol{v}) = \phi_f(\mathbf{f}_{\text{G}}(\boldsymbol{x}), \boldsymbol{v})$ of $G_{\mathbf{0}}$, respectively. $\mathbf{f}$ could be directly transformed to color $\boldsymbol{c}(\boldsymbol{x}, \boldsymbol{v}) = \phi_c(\mathbf{f}(\boldsymbol{x}, \boldsymbol{v}))$ or being volume integrated to $\mathbf{F}$ and sent to $G_{\mathbf{1}}$ for high resolution synthesis. For simplicity, we will omit $\boldsymbol{v}$ in the following.

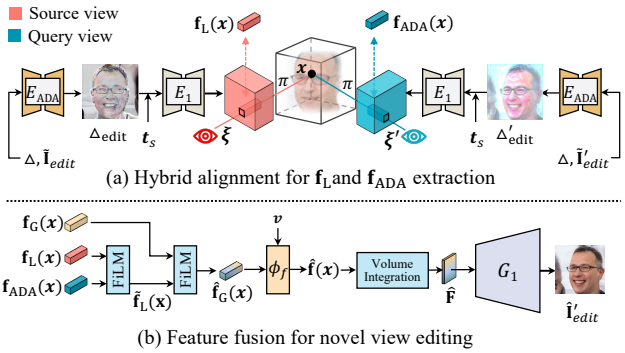
Local Feature for Detailed Textures. The global latent code $\hat{\mathbf{w}}$ is a compact representation of the predicted scene. However, previous works [9, 49] have validated that a low-dimensional latent code discards high-frequency spatial details and fails to reconstruct high-fidelity outputs. This phenomenon becomes more severe when lifting the 2D image



(a) Self-supervised Inversion Learning

(b) Local Feature Fusion for High-Fidelity Inversion

Figure 2. **E3DGE for 3D GAN inversion**. (a) We augment the training of the encoder $E_{\mathbf{0}}$ with 3D supervision $\mathcal{L}_{geo}$ for plausible 3D shape prediction. (b) We augment the representation capacity of the global latent code $\hat{\mathbf{w}}$ with local point-dependent latent feature $\mathbf{f}_{\text{L}}$ for high-fidelity texture reconstruction.



(a) Hybrid alignment for $\mathbf{f}_{\text{L}}$and $\mathbf{f}_{\text{ADA}}$ extraction

(b) Feature fusion for novel view editing

Figure 3. **Hybrid alignment for high-quality editing.** Given code prediction $\hat{\mathbf{w}}$ from encoder $E_{\mathbf{0}}$ pre-trained in stage-I, we aim to generate high-quality view synthesis over the edited code $\hat{\mathbf{w}}_{\text{edit}}$. In (a), the local details Δ along with the target edited image $\mathbf{I}'_{\text{edit}}$ and depth map $\boldsymbol{t}_s(\hat{\mathbf{w}}, \boldsymbol{\xi})$ are sent to pre-trained E_{ADA} to predict aligned residual Δ'_{edit}. The original aligned residual Δ along with the 2D auxiliary residual Δ'_{edit} are processed by $E_{\mathbf{1}}$ to recover latent maps $\mathbf{F}_{\text{L}}$ and $\mathbf{F}_{\text{ADA}}$ for later fusion. In (b), the extracted features $\mathbf{f}_{\text{L}}(\boldsymbol{x})$ and $\mathbf{f}_{\text{ADA}}(\boldsymbol{x})$ are first fused together with a FiLM layer, and the fused result $\tilde{\mathbf{f}}_{\text{L}}(\boldsymbol{x})$ further serve as conditions to modulate the global feature $\mathbf{f}_{\text{G}}(\boldsymbol{x})$. The final modulated feature $\hat{\mathbf{f}}(\boldsymbol{x})$ contains complete information, globally and locally. The volume integrated $\hat{\mathbf{F}}$ is sent to $G_{\mathbf{1}}$ for high-resolution synthesis.

to a 3D scene, which contains exponentially more information. Inspired by recent progress in few-shot 3D reconstruction [3, 10, 42, 43, 48, 51, 54], we propose to make up for the lost information by introducing pixel-aligned (local) features. As shown in Fig. 2 (b), rather than conditioning all 3D points with the same latent code $\hat{\mathbf{w}}$, we augment the representation capacity with local latent codes $\mathbf{f}_{\text{L}}$ that is depen-

dent on each point $\boldsymbol{x}$. We introduce a local hourglass [30] encoder $E_{\mathbf{1}}$ to predict a residual feature map $\mathbf{F}_{\mathrm{L}}$ based on the reconstruction residue $\Delta = \mathbf{I} - \tilde{\mathbf{I}}$,

$$\mathbf{F}_{\mathrm{L}} = E_{\mathbf{1}}(\Delta, \boldsymbol{t}_s(\hat{\mathbf{w}}, \boldsymbol{\xi})), \tag{7}$$

where $\boldsymbol{t}_s(\hat{\mathbf{w}}, \boldsymbol{\xi})$ is the depth map of the scene derived from the SDF to serve as 3D context information. Then, the local latent code of a point $\boldsymbol{x}$ is its corresponding value in $\mathbf{F}_{\mathrm{L}}$:

$$\mathbf{f}_{\mathrm{L}}(\boldsymbol{x}) = \mathbf{F}_{\mathrm{L}}(\pi(\boldsymbol{x})) \oplus \mathbf{PE}(\boldsymbol{x}), \tag{8}$$

where π maps the 3D point $\boldsymbol{x}$ to its corresponding pixel coordinate on 2D feature map $\mathbf{F}_{\mathrm{L}}$. Since in 3D scenes, points along a ray will be projected to the same coordinate on the 2D plane, to differentiate these points, we additionally concatenate their positional encoding $\mathbf{PE}(\boldsymbol{x})$ [29] in Eq. (8). In this way, the local feature $\mathbf{f}_{\mathrm{L}}$ only encodes the residual information at the projected position $\pi(\boldsymbol{x})$ but is also capable of determining where the residual information lies in the 3D scene, as well as inpainting the occluded areas along the ray.

Finally, we fuse the local latent code $\mathbf{f}_{\mathrm{L}}(\boldsymbol{x})$ with the global latent code $\mathbf{f}_{\mathrm{G}}(\boldsymbol{x}) = E_{G_{\mathbf{0}}}(\boldsymbol{x}, \hat{\mathbf{w}})$ to supplement the missing high-frequency details. Specifically, the feature fusion is based on Feature-wise Linear Modulation (FiLM) [38]. As shown in Fig. 2, $\mathbf{f}_{\mathrm{L}}(\boldsymbol{x})$ is fed into two MLP layers to obtain the scale and bias modulation parameters $\mathbf{f}_{\mathrm{L}}^{\gamma}(\boldsymbol{x})$ and $\mathbf{f}_{\mathrm{L}}^{\beta}(\boldsymbol{x})$. Then we modulate $\mathbf{f}_{\mathrm{G}}(\boldsymbol{x})$ with FiLM

$$\hat{\mathbf{f}}_{\mathrm{G}}(\boldsymbol{x}) = \mathrm{FiLM}(\mathbf{f}_{\mathrm{G}}(\boldsymbol{x}), \mathbf{f}_{\mathrm{L}}(\boldsymbol{x})) = \mathbf{f}_{\mathrm{L}}^{\gamma}(\boldsymbol{x}) \cdot \mathbf{f}_{\mathrm{G}}(\boldsymbol{x}) + \mathbf{f}_{\mathrm{L}}^{\beta}(\boldsymbol{x}).$$

The fused $\hat{\mathbf{f}}_{\mathrm{G}}(\boldsymbol{x})$ is volume integrated to $\hat{\mathbf{F}}$ and the final high-fidelity reconstructed image is obtained as $\hat{\mathbf{I}} = G_{\mathbf{1}}(\hat{\mathbf{F}})$.

Note that through point projection π, the reconstruction with local prior is not limited to the original view, and naturally works for novel views. However, for views with severe occlusions or additional editing, the residual features may not fully align with the scene, leading to a failed feature fusion. We will address this issue in the next subsection with our hybrid feature alignment.

4.3. Hybrid Alignment for High-Quality Editing

Though we achieve high-fidelity reconstruction with the aforementioned designs, there is a trade-off between the *input view* reconstruction quality and *novel view* editing performance. We first analyze the reasons behind and propose a hybrid alignment module to address this issue.

Reconstruction Editing Trade-off. Given an input image $\mathbf{I}$ with paired reconstruction $\tilde{\mathbf{I}}$ and residual map Δ extracted from the input view $\boldsymbol{\xi}$ with the aforementioned method, the reconstruction performance trade-offs the editing performance due to the following two reasons. First, at test time when the input image is edited $\tilde{\mathbf{I}}_{\text{edit}}$ or query view $\boldsymbol{\xi}' \neq \boldsymbol{\xi}$, the residual map no longer aligns and is likely to result in wrong predictions. Second, if we supervise the models to reconstruct the input itself, the learned features are *regressive* rather than *generative* since all prediction areas are visible in the inputs. With these above-mentioned challenges, though the model could yield perfect reconstruction at training, it would result in noticeable performance degradation when rendering from novel views at test time.

Hybrid Alignment for High-Quality Editing. To address the first challenge, we propose to infer aligned features with a 2D-3D hybrid alignment. Specifically, given edited latent code $\hat{\mathbf{w}}_{\text{edit}}$, the initial novel-view edited image $\tilde{\mathbf{I}}'_{\text{edit}} = G_{\mathbf{0}}(\hat{\mathbf{w}}_{\text{edit}}, \boldsymbol{\xi}')$ is misaligned with Δ. Inspired by HFGI [49], we leverage a 2D alignment module E_{ADA} to address the misalignment. As shown in Fig. 3 (a), we first obtain $\Delta_{\text{edit}} = E_{\text{ADA}}(\Delta, G_{\mathbf{0}}(\hat{\mathbf{w}}_{\text{edit}}, \boldsymbol{\xi}))$, transform it to residual feature map $\mathbf{F}_{\mathrm{L}}^{\text{edit}}$ via Eq. (7) and retrieve the view-consistent 3D local feature $\mathbf{f}_{\mathrm{L}}$ via Eq. (8). However, to render the high-quality edited image $\hat{\mathbf{I}}'_{edit}$ from novel view $\boldsymbol{\xi}'$, $\mathbf{F}_{\mathrm{L}}^{\text{edit}}$ might still suffer from occlusion due to large pose variations. To the end, we propose a hybrid alignment to further refine $\mathbf{F}_{\mathrm{L}}^{\text{edit}}$ with 2D aligned feature from E_{ADA}. Specifically, we align a 2D residue $\Delta'_{\text{edit}} = E_{\text{ADA}}(\Delta, \tilde{\mathbf{I}}'_{\text{edit}})$ and retrieve its corresponding $\mathbf{f}_{\text{ADA}}$ with $E_{\mathbf{1}}$, which fills the occlusion in a 2D manner but lacks 3D consistency. To marry the best of both, as shown in in Fig 3 (b), we modulate $\mathbf{f}_{\mathrm{L}}$ with $\mathbf{f}_{\text{ADA}}$,

$$\tilde{\mathbf{f}}_{\mathrm{L}}(\boldsymbol{x}) = \mathrm{FiLM}(\mathbf{f}_{\mathrm{L}}(\boldsymbol{x}), \mathbf{f}_{\text{ADA}}(\boldsymbol{x})), \tag{9}$$

and further fuse $\tilde{\mathbf{f}}_{\mathrm{L}}$ with $\mathbf{f}_{\mathrm{G}}(\boldsymbol{x})$ for final prediction,

$$\hat{\mathbf{f}}(\boldsymbol{x}) = \mathrm{FiLM}(\mathbf{f}_{\mathrm{G}}(\boldsymbol{x}), \tilde{\mathbf{f}}_{\mathrm{L}}(\boldsymbol{x})), \tag{10}$$

where $\hat{\mathbf{f}}(\boldsymbol{x})$ is then integrated to $\hat{\mathbf{F}}$ for rendering the final novel-view edited image $\hat{\mathbf{I}}'_{edit} = G_{\mathbf{1}}(\hat{\mathbf{F}})$.

Novel View Training for Coherent View Synthesis. To address the second challenge and enforce the model to learn generative features, during training, we sample two views $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ for each style code $\mathbf{w}$, and render the corresponding images $\mathbf{I}^{\boldsymbol{\xi}_1}$ and $\mathbf{I}^{\boldsymbol{\xi}_2}$. Then, we train the models to reconstruct plausible novel views, *i.e.*, $G(E(\mathbf{I}^{\boldsymbol{\xi}_1}), \boldsymbol{\xi}_2) \approx \mathbf{I}^{\boldsymbol{\xi}_2}$ and $G(E(\mathbf{I}^{\boldsymbol{\xi}_2}), \boldsymbol{\xi}_1) \approx \mathbf{I}^{\boldsymbol{\xi}_1}$. This training strategy facilitates a high-quality view synthesis over edited scenes.

Training. We leverage the image reconstruction loss [1], defined as $\mathcal{L}_{rec}(\mathbf{I}) = \lambda_1 \mathcal{L}_2(\mathbf{I}) + \lambda_2 \mathcal{L}_{\text{LPIPS}}(\mathbf{I}) + \lambda_3 \mathcal{L}_{\text{Id}}(\mathbf{I})$. We further adopt adversarial loss $\mathcal{L}_{\text{adv}}$ [21] to improve the naturalness of the output image.

5. Experiments

Datasets. We mainly focus on the human face domain and use both 2D and 3D datasets for extensive evaluation. To examine 2D reconstruction quality, we adopt CelebA-HQ [20, 24] dataset for source view reconstruction. To further evaluate novel view synthesis performance, we synthesize 100 trajectory videos from a pretrained generator as

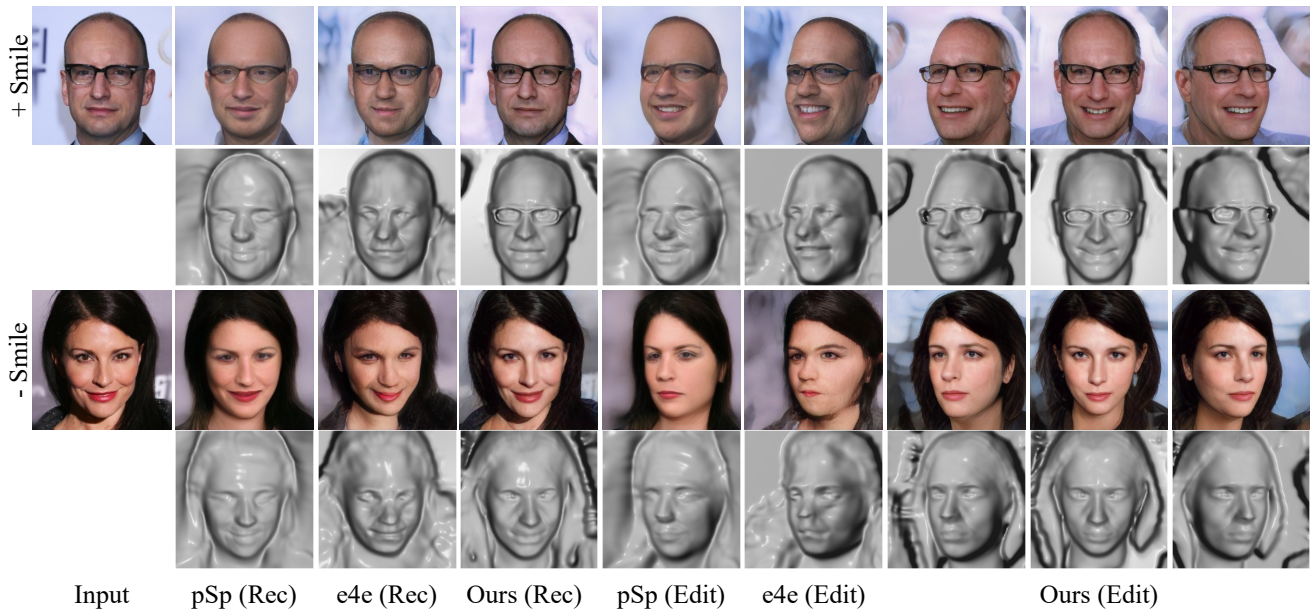


Figure 4. **Qualitative comparisons on face reconstruction (Rec) and editing (Edit) under novel views.**

Table 1. **Quantitative performance on CelebA-HQ.** 'T' and 'S' denote the time for texture and shape inversion, respectively.

Methods	MAE ↓	SSIM ↑	LPIPS ↓	Similarity ↑	Time(s) ↓
$SG2_{EG3D}$	.241 ± .019	.671 ± .014	.288 ± .019	.434 ± .037	100
PTI_{EG3D}	.079 ± .005	.769 ± .012	.105 ± .011	.779 ± .027	114
$SG2_{StyleSDF}$	.202 ± .063	.650 ± .054	.167 ± .046	.219 ± .106	235
$PTI_{StyleSDF}$	**.062 ± .012**	**.796 ± .017**	**.027 ± .005**	**.892 ± .009**	246
$pSp_{StyleSDF}$	.150 ± .032	.696 ± .048	.270 ± .059	.498 ± .099	0.29
$e4e_{StyleSDF}$	.174 ± .049	.669 ± .049	.226 ± .063	.252 ± .107	0.29
E3DGE	**.103 ± .010**	**.769 ± .039**	**.136 ± .039**	**.881 ± .041**	**0.45(T)/0.81(S)**

Table 2. **Quantitative performance on Novel View Synthesis.**

Methods	MAE ↓	SSIM ↑	LPIPS ↓	Similarity ↑
$SG2_{StyleSDF}$	.284 ± .025	.572 ± .006	.244 ± .031	.304 ± .036
$PTI_{StyleSDF}$	.186 ±.016	.652 ± .015	.215 ± .045	.795 ± .040
$pSp_{StyleSDF}$	.201 ± .010	.634 ± .005	.285 ± .029	.559 ± .043
$e4e_{StyleSDF}$	.197 ± .016	.597 ± .011	.212 ± .023	.297 ± .058
E3DGE	**.147 ± .011**	**.694 ± .018**	**.151 ± .024**	**.901 ± .012**

a proxy test set. For attribute editing, we adopt InterfaceGAN [46] and Talk2Edit [19] to search for the editing directions. To evaluate 3D shape reconstruction quality, we use NoW benchmark [44] that provides a rich variety of face images with ground-truth 3D scans. The 3D GANs are pre-trained on FFHQ [21]. Note that our method does not rely on any external 3D data during the training process.

Implementation Details. For all the encoder models, we adopt Adam optimizer with a learning rate of $5e-5$ to train the models on 4 NVIDIA Tesla V100 GPUs, with a resolution of 256^2, batch size of 24, and 16 samples along a ray for the recommended $200K$ iterations. Following [42], we filter our invisible 3D points when training from a certain view. Code, dataset, and all pre-trained models will be made publicly available. More details are included in the supplementary material.

5.1. Evaluation

5.1.1 Quantitative Evaluation

For comparison, we implement two canonical encoder-based GAN inversion approaches on StyleSDF [33], *i.e.*, pSp [40] and e4e [47], which stress reconstruction and editing quality respectively. Furthermore, we also implement optimization-based methods [21,41] on StyleSDF and EG3D [8] for extensive comparison.

We report inversion performance for both source view reconstruction and novel view synthesis in Tabs 1-2. For source view reconstruction, the metrics are calculated on the $2,824$ images from CelebA-HQ test set [24]. For novel view synthesis, the metrics are averaged from 100 videos generated from pre-trained 3D GANs, each with 250 frames covering ellipsoid camera poses trajectory. For each video, we randomly pick one image as source view input and the remaining images as ground truths with labeled poses as query views. In this way, we could extensively evaluate the view synthesis ability under occlusions and varied input viewpoints. We also compare E3DGE against two optimization-based methods [22,41]. As demonstrated in Tab 1, our approach substantially outperforms encoder-based baselines in terms of reconstruction quality on two settings and achieves considerablely faster inference speed against optimization-based methods. Notice that we do not include EG3D in Tab 2 due to its camera pose being mis-

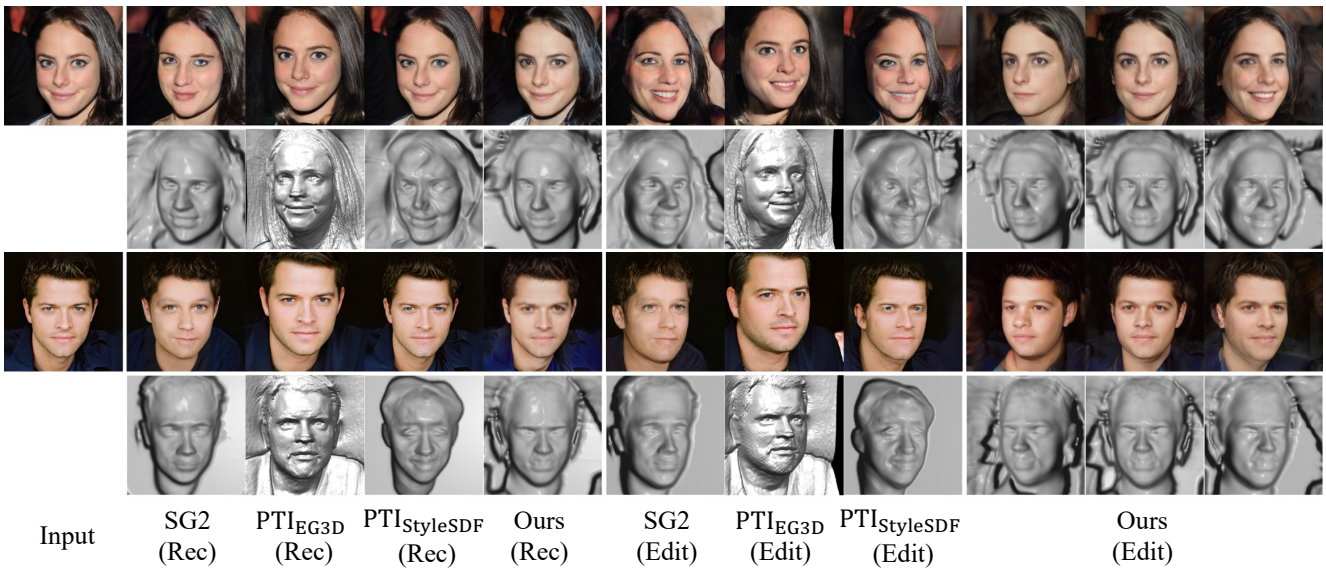


Figure 5. Visual comparisons on optimization-based methods. 'Rec' and 'Edit' represent reconstruction and editing, respectively.

aligned with StyleSDF. Please also refer to the supplementary materials for quantitative results of 3D reconstruction.

5.1.2 Qualitative Evaluation

Encoder Baselines. We visualize both inversion and editing results against encoder baselines in Fig. 4. *Geometry-wise*, the baseline models without explicit 3D supervisions tend to generate implausible intermediate shapes, *e.g.*, Besides, their reconstruction is not close to the "ground truth", and the reconstructed surface lacks details. Our method successfully regularizes the intermediate 3D shapes and generates plausible results with surface details and a more complete structure. For instance, in rows 1, our method reconstructs 3D eyeglasses in which the baselines fail. Corresponding metrics in Tab. 4 also validate the usefulness of the direct geometry supervisions and loss designs. *Texture-wise*, existing methods generate distorted results and suffer artifacts and identity change. In contrast, with pixel-aligned features incorporated, our method is more robust with high-fidelity results. In particular, our method captures more details and preserves the identity of different input viewpoints.

For editing, we choose the "Smile" attribute for editing. Beyond plausible shape reconstruction with high-fidelity texture inversion, and in-view synthesis over edited results, our method consistently generates high-quality edited renderings in terms of view consistency, details conservation, and identity preservation. Compared with our method, the baselines either fail to render intact identity (column 5) or generate visually plausible shapes (column 6).

Optimization Baselines. We also compare our method with the state-of-the-art optimization-based methods [8, 41] in Fig. 5. We include the performance of PTI [41] on both StyleSDF and EG3D for extensive evaluation. With more than 100× faster inference, our method achieves a comparable inversion quality. Also, the editing results produced by E3DGE successfully preserve the local details with high-fidelity novel view editing performance. We also notice the geometry-texture misalignment of EG3D (column 7), where the "Smiling" texture with teeth does not align with the geometry without teeth.

5.2. Ablation Study

Effect of 3D GAN as Supervisions. We quantitatively validate the effects of 3D supervision in the NoW Challenge validation set and report the corresponding metrics in Tab. 4. Compared with 2D supervision only, adding 3D supervisions greatly improves the reconstruction quality. We also validate the benefits of all loss terms in E_0 training.

Effect of Local Features. As discussed, the local features preserve the image details to facilitate high-fidelity reconstruction. To validate the effectiveness of local features in texture reconstructions, we show the inversion results in Fig. 6. With the proposed local-global fusion pipeline, our model captures more details and guarantees photorealistic reconstruction. Quantitative results in Tab. 3 also validate the effectiveness of local features in high-quality inversion. The results on the video trajectories also show that without delicate design, *e.g.* novel view training, local features would fully collapse over novel view synthesis.

Effect of Hybrid Alignment. We show the view synthesis achieved by different alignment methods in Fig. 7. To quantitatively analyze the effect of hybrid alignment, in Tab. 3 we evaluate the model performance of 3D alignment and 2D alignment individually. For both ablations, novel view training is enabled. As shown here, the 3D alignment model shows better view consistency in video prediction

Table 3. **Ablations of Local Features and Hybrid Fusion.** Our local-global model design with hybrid alignment achieves the balance of high-quality reconstruction and view synthesis.

	Source View Reconstruction				Novel View Synthesis			
Ablation Settings	MAE ↓	SSIM ↑	LPIPS ↓	ID ↑	MAE ↓	SSIM ↑	LPIPS ↓	ID ↑
Synthetic Training	.245 ± .024	.634 ± .019	.333 ± .029	.369 ± .056	.241 ± .011	.594 ± .008	.366 ± .059	.770 ± .026
+Local Features	**.074 ± .007**	**.811 ± .015**	**.075 ± .010**	**.953 ± .006**	.282 ± .103	.571 ± 0.056	.511 ± 0.031	.608 ± .123
+2D Alignment	.098 ± .005	.774 ± .038	.140 ± .040	.900 ± .032	.178 ± .007	.656 ± .009	.178 ± .012	**.895 ± .018**
+3D Alignment	.102 ± .009	.772 ± .015	.119 ± .016	.818 ± .029	**.150 ± .011**	**.689 ± .022**	**.140 ± .021**	.891 ± .011

Table 4. Effect of 3D Supervisions on the NoW Challenge.

Settings	Median↓	Mean↓	Std
$\text{pSp}_{\text{StyleSDF}}$	1.97	2.43	2.05
$\text{e4e}_{\text{StyleSDF}}$	2.83	3.40	2.67
$+\mathcal{L}_{geo}^{\mathcal{O}}$	1.75	2.11	1.72
$+\mathcal{L}_{geo}^{\mathcal{F}}$	1.71	2.09	1.70
$+\mathcal{L}_{code}$	**1.66**	**2.06**	**1.69**

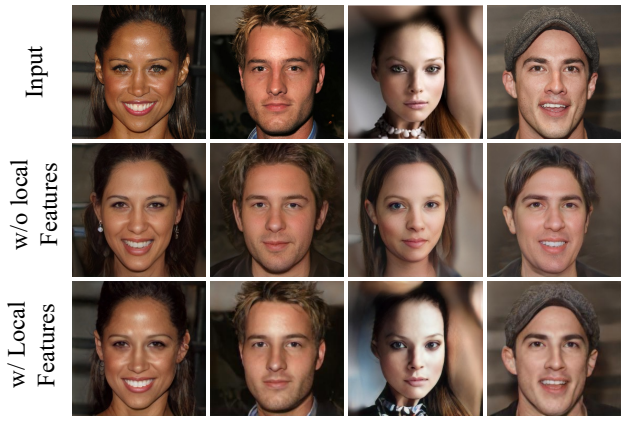


Figure 6. **Ablation of Local Features.** Our method with pixel-aligned features shows photorealistic reconstructions.

measured by reconstruction metrics, and the 2D alignment model shows better identity preservation. The hybrid alignment model marries the best of both and also enables semantic editing and yields better reconstruction performance on the video predictions.

6. Conclusion and Discussions

We propose a novel 3D GAN inversion framework E3DGE for 3D GAN inversion and editing. We marry the benefits of both self-supervised global prior and pixel-aligned local prior for high-quality shape and texture reconstruction. A hybrid alignment that bridges the best of 2D and 3D features is further proposed for view-consistent editing. Benefiting from the overall system design, the proposed method has advantages in terms of both high fidelity and editability. As a pioneer attempt in this direction, we believe this work opens a new line of research direction and will inspire future works on 3D GAN inversion, few-shot 3D reconstruction and 3D-aware learning from 2D images.

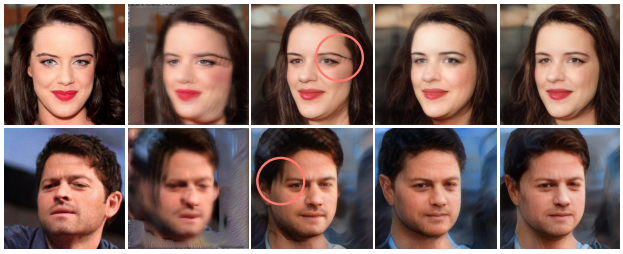


Figure 7. **Ablation of Hybrid Alignment.** From left to right, we show the novel view synthesis of raw 3D-aligned features w/wo novel view training, synthesis achieved using 2D-aligned features, and the final hybrid features. 3D-aligned features are view-consistent but suffer from occlusions (circled), while 2D features are visually plausible but lack some details (*e.g.*, hair color). Our hybrid fused results share the best of both.

Limitations and Future Work. The proposed method suffers data bias introduced by the synthetic data. As the synthetic data lacks complex details and pose variations compared with real-world data, our method trained with it tends to generate simple background and fail on extreme poses. Special attentions should be paid to data bias to avoid social impact to under represented minorities. A future direction is to leverage real data for semi-supervised training. Another future direction is to leverage the hyper-network for efficient local feature incorporation to alleviate the extra computational cost of the 2D alignment module. Finally, we would explore the potentials of our framework on other 3D GANs and shapes beyond human face and other editing methods uniquely designed for 3D GANs.

Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). It is also partially supported by Singapore MOE AcRF Tier 2 (MOE-T2EP20221-0011) and the NTU URECA research program.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 1, 2, 5

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN++: How to edit the embedded images? In *CVPR*, 2020. 1, 2

[3] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3D reconstruction of humans wearing clothing. *CVPR*, 2022. 4

[4] Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick P'erez. This Dataset Does Not Exist: Training Models from Generated Images. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. 2

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*. OpenReview.net, 2019. 2

[6] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2NeRF: Unsupervised Conditional p-GAN for Single Image to Neural Radiance Fields Translation. In *CVPR*, 2022. 2

[7] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and G. Wetzstein. Pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In *CVPR*, 2021. 2, 3

[8] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 2, 3, 6, 7

[9] Kelvin C.K. Chan, Xiangyu Xu, Xintao Wang, Jinwei Gu, and Chen Change Loy. GLEAN: Generative latent bank for large-factor image super-resolution and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 4

[10] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo Radiance Fields (SRF): Learning View Synthesis for Sparse Views of Novel Scenes. *CVPR*, 2021. 4

[11] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. In *SIGGRAPH*, volume 40, 2021. 2

[12] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *ECCV*, 2022. 2

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[14] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. StyleNeRF: A style-based 3D-aware generator for high-resolution image synthesis. In *ICLR*, 2021. 2, 3

[15] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3D shape from adversarial rendering. In *ICCV*, 2019. 2

[16] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 2

[17] Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. *The International Conference on Learning Representations (ICLR)*, 2020. 2

[18] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *ICLR*, 2022. 2

[19] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-Edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 6

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 5

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 3, 5, 6

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 2, 6

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 3

[24] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 5, 6

[25] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *TOG*, 36(6), 2017. 2

[26] Connor Z. Lin, David B. Lindell, Eric Chan, and Gordon Wetzstein. 3D GAN Inversion for Controllable Portrait Image Animation. *arXiv*, abs/2203.13441, 2022. 2

[27] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. In *NIPS*, 2021. 2

[28] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *CVPR*, June 2019. 2

[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 *ECCV*. Springer, 2020. 2, 3, 5

[30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 5

[31] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yongliang Yang. HoloGAN: Unsupervised Learning of 3D Representations From Natural Images. In *ICCV*, 2019. 2

[32] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 2, 3

[33] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *CVPR*, 2021. 2, 3, 6

[34] 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. In *ICLR*, 2021. 2

[35] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A Shading-Guided Generative Implicit Model for Shape-Accurate 3D-Aware Image Synthesis. In *NIPS*, 2021. 3

[36] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting Deep Generative Prior for Versatile Image Restoration and Manipulation. *TPAMI*, 44:7474–7489, 2022. 2

[37] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*. IEEE, 2019. 2, 4

[38] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 5

[39] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look, 2022. 2

[40] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a StyleGAN encoder for image-to-image translation. In *CVPR*, 2021. 1, 2, 3, 4, 6

[41] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 1, 2, 6, 7

[42] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, October 2019. 4, 6

[43] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR*, June 2020. 4

[44] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[45] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative radiance fields for 3D-aware image synthesis. In *NIPS*, 2020. 2

[46] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterFaceGAN: Interpreting the disentangled face representation learned by GANs. *PAMI*, PP, 2020. 6

[47] 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, 6

[48] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser. IBRNet: Learning Multi-View Image-Based Rendering. In *CVPR*, pages 4688–4697, 2021. 4

[49] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-Fidelity GAN inversion for image attribute editing. In *CVPR*, 2022. 1, 2, 4, 5

[50] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. GAN Inversion: A Survey. *TPAMI*, 2022. 1

[51] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *CVPR*, June 2022. 4

[52] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18409–18418, 2021. 2

[53] Shuai Yang, Liming Jiang, Ziwei Liu, , and Chen Change Loy. VToonify: Controllable high-resolution portrait video style transfer. *ACM Transactions on Graphics (TOG)*, 41(6):1–15, 2022. 2

[54] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. PixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 4

[55] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. DatasetGAN: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021. 2

[56] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain GAN Inversion for Real Image Editing. In *ECCV*, 2020. 2