GeoGen: Geometry-Aware Generative Modeling via Signed Distance Functions

Salvatore Esposito[1]† Qingshan Xu[3] Kacper Kania[4] Charlie Hewitt[2]
Octave Mariotti[1] Lohit Petikam[2] Julien Valentin[2] Arno Onken[1]
Oisin Mac Aodha[1]

[1]University of Edinburgh [2]Microsoft
[3]Huazhong University of Science and Technology [4]Warsaw University of Technology

Abstract

We introduce a new generative approach for synthesizing 3D geometry and images from single-view collections. Most existing approaches predict volumetric density to render multi-view consistent images. By employing volumetric rendering using neural radiance fields, they inherit a key limitation: the generated geometry is noisy and unconstrained, limiting the quality and utility of the output meshes. To address this issue, we propose GeoGen, a new SDF-based 3D generative model trained in an end-to-end manner. Initially, we reinterpret the volumetric density as a Signed Distance Function (SDF). This allows us to introduce useful priors to generate valid meshes. However, those priors prevent the generative model from learning details, limiting the applicability of the method to real-world scenarios. To alleviate that problem, we make the transformation learnable and constrain the rendered depth map to be consistent with the zero-level set of the SDF. Through the lens of adversarial training, we encourage the network to produce higher fidelity details on the output meshes. For evaluation, we introduce a synthetic dataset of human avatars captured from 360-degree camera angles, to overcome the challenges presented by real-world datasets, which often lack 3D consistency and do not cover all camera angles. Our experiments on multiple datasets show that GeoGen produces visually and quantitatively better geometry than the previous generative models based on neural radiance fields.

1. Introduction

The combination of generative models [19–21, 26] and implicit neural representations [7, 25, 32] has sparked considerable advancements in 3D representation learning [4, 14]. It has powered the synthesis of high-quality, multi-view consistent, images. However, a common pitfall in the pursuit of higher image quality is the sidelining of the quality of the underlying *geometry* [43].

Recent non-generative efforts, such as NeuS [39], VolSDF [43], and Geo-Neus [11], have made use of the zero-level set of a Signed Distance Function (SDF) to represent the surface of the geometry in a scene via a surface rendering equation, ultimately achieving high-fidelity scene reconstruction. While these models have shown impressive potential, given their non-generative nature, they are only able to reconstruct a scene of interest when multi-view image data is available. This limitation highlights the need for generative models capable of producing high-quality 2D images that are suitable for content creation while ensuring precise geometric synthesis without multi-view data.

Other recent methods such as Ball-GAN [36], and EG3D [5], have combined generative models with Neural Radiance Fields (NeRFs) [26] to yield high quality rendered images. Yet, these approaches often result in noisy meshes that contain geometric artifacts, which emerge due to the properties of NeRFs and their lack of constraints on the geometric reconstructions. Attempts have also been made to harmonize SDFs with generative models as in [31]. However, the generated meshes are often overly smooth, a result of the smoothing prior that encourages the SDF to produce valid values everywhere in 3D space. Additionally, applying this loss can be prohibitive at higher resolutions.

In this work, we address these issues by adding SDF constraints to improve the synthesized geometry of a 3D-aware generative model. Our approach, named GeoGen, employs an SDF depth map consistency loss for enhanced geometric generation. Specifically, we build on EG3D [5] by introducing an SDF representation, instead of a density representation, to encode the geometry. This allows GeoGen to extract mesh surfaces directly from the zero-level set of the SDF [30, 39, 43]. In order to make the SDF representation learning feasible, and to endow it with the ability to model

†Work conducted during an internship at Microsoft.

complex and detailed geometry, we also propose an SDF depth map consistency loss. We use a fixed density-to-SDF transformation function to convert the density representation to an SDF. This facilitates generative feature learning by making the learning objective easier to optimize. The SDF also enables the extraction of smooth depth maps that serve as a 'pseudo' ground-truth. Our approach uses its own depth prediction in a self-supervised manner to improve the reconstruction. In contrast to commonly used priors, our approach is cheap to compute with only a minor increase in training time.

GeoGen is able to generate detailed meshes from a single input 2D image via inversion [33]. This capability is valuable in applications where the requirement for detailed and realistic meshes is needed. In stark contrast to recent methods like Rodin [40], which required 30 million images during training to create 3D meshes, GeoGen uses a fraction of this number – approximately 50,000 images. Other methods such as PanoHead [1] propose an augmented triplane and separate foreground and background in 2D images with the help of a custom in-house dataset. However, with our proposed architecture, we show that by enforcing our geometric constraints, we are able to reconstruct a detailed 360° geometry, with a reduction in visual artifacts (*e.g.* the backs of heads) compared to methods such as EG3D [1].

We make the following contributions: (i) We address the problem of 3D synthesis from 2D images by combining a Signed Distance Function (SDF) network with a StyleGAN generative architecture. Our GeoGen model produces more refined geometry predictions compared to conventional neural volume rendering. (ii) We propose an SDF depth map consistency loss that is designed to address geometric inaccuracies from volumetric integration by aligning 3D points with the SDF network's zero-level set for more precise reconstructions. (iii) We introduce a new dataset of realistic synthetic human heads that contains 360° camera views from multiple synthetic humans. This dataset will be a valuable resource for training and quantitatively evaluating 3D generative models. It can be found on our webpage https://microsoft.github.io/GeoGen.

2. Related work

The landscape of generative modeling has seen a shift in recent years, with techniques drawing on neural implicit representations, such as Generative Adversarial Networks (GANs) [14] and Diffusion models [10, 18, 24, 37] emerging as powerful tools. These techniques blend generative models with neural volume rendering, thereby synthesizing 3D images that capture novel viewpoints from 2D data alone [26]. However, a recurring challenge in this domain has been the reliance on generic density functions to learn the geometry of the images, a factor that often introduces artifacts and results in noisy, low-quality geometric predictions [31]. To mitigate this, prior work has taken advantage of large amounts of multi-view data to constrain the models, thereby yielding more robust geometry [39, 43], but at the expense of not being fully generative.

The emergence of volumetric implicit representations, bolstered by the strengths of Multi-Layer Perceptrons (MLPs) [15] and neural rendering techniques [26], has shown substantial promise in extracting detailed geometry from a 3D scene. This is most apparent in methods such as NeuS [39] and VolSDF [43], which extract high-fidelity surfaces by representing the scene using the Signed Distance Function (SDF) and extracting the surface at the zero level set.

Meanwhile, the broader field of deep learning has seen a surge in novel methods for creating 3D representations from 2D data. One such family of methods is Neural Radiance Fields (NeRFs) [26], which employs a neural network to model the radiance of a 3D scene at any spatial point. The ability of NeRFs to generate high-fidelity 3D models from 2D multi-view supervision, complete with accurate lighting and shading effects, makes them an attractive option for applications requiring realistic 3D representations, such as virtual reality [43].

One set of methods that deserves particular discussion within this landscape is the set of 3D-aware generative models [4, 9, 12, 13, 16, 27–29, 35]. These methods are specifically designed to generate 3D representations of objects or scenes, utilizing a variety of techniques, including volumetric representations, SDFs, and implicit neural representations. For instance, the Generative Radiance Fields (GRAF) model [34] generates high-resolution 3D shapes with intricate detail, leveraging a neural network to model the radiance and shape of a 3D object. Other notable models include DeepSDF [32], which learns continuous signed distance functions for arbitrary shapes using 3D supervision, and HoloGAN [27], which generates 3D objects by imposing structural constraints in the generative process. Recently, EG3D [5] proposed a triplane representation for volume rendering in generative models, which enables efficient 3D-aware generation. However, extracting high quality 3D meshes is not guaranteed because of its use of a volume density representation. StyleSDF [31], makes use of an SDF representation to directly model geometry, but the extracted surfaces are overly smooth making it challenging to use them in practical applications.

In our investigation of 3D-aware generative models and SDF representations, we identify certain limitations inherent in existing methodologies. One such limitation appears to be a result of the use of the Eikonal loss [11, 39, 43], leading to overly smooth geometry synthesis. Our methodology, building on the foundation laid by EG3D, aims to overcome this by introducing an SDF depth-consistency constraint. This novel constraint is designed to refine geometric surface

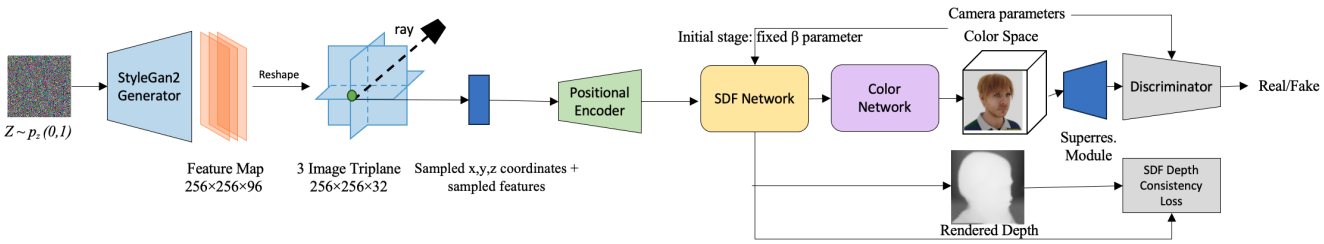

Figure 1. GeoGen, our 3D-aware generator, is trained solely from 2D images. Noise sampling is followed by a StyleGan2 generator that produces triplane features similar to EG3D [5]. However, we enhance them with positional info and an SDF network for refined geometry. GeoGen is end-to-end trained with a GAN objective along with our SDF depth consistency loss.

predictions by leveraging a self-supervised depth prediction mechanism. Unlike previous efforts, such as StyleSDF [31], which merely translates SDF values into density fields, our approach harnesses the full potential of SDF for geometry representation as exemplified by VolSDF [43]. We emphasize that incorporating our SDF representation and its associated constraints does not substantially complicate the training of generative models yet provides enhanced control over geometric surface detail.

3. Method

Here we present our GeoGen generative approach for enhanced geometric synthesis. We begin by revisiting EG3D [5], an efficient geometry-aware 3D GAN that introduces notation and provides context for our contributions. Then we describe our SDF-based generative model which builds on the EG3D framework.

3.1. Efficient geometry-aware 3D GAN

EG3D [5] is an efficient geometry-aware 3D generative adversarial network. It consists of a StyleGAN2 [20] based feature generator, triplane representation, implicit volume render, and super-resolution module. In order to generate an image, it first samples a random latent noise code and processes the code via a mapping network. The processed code is used to drive the StyleGAN2 generator to produce feature maps which are reshaped to form three feature planes. During the volume rendering, a queried 3D point $\boldsymbol{p}$ is projected onto each of the three feature planes, leading to corresponding feature vector $[F_{xy}(\boldsymbol{p}), F_{xz}(\boldsymbol{p}), F_{yz}(\boldsymbol{p})]$. These feature vectors are further processed by a shallow MLP to yield the color and density at the position $\boldsymbol{p}$. By the process of volumetric integration, a low-resolution image is generated based on the sampled points along all image rays. Finally, a super-resolution module is used to generate high-resolution output images.

Like EG3D, we also use a triplane representation to efficiently generate images. Different from EG3D, which targets geometry-aware *image* synthesis, we focus on high-quality *geometry* synthesis. To this aim, we introduce an SDF-based generative model and present a novel SDF learning strategy.

3.2. SDF-based generative model

Our goal is to develop a model that can learn to generate 3D consistent object-centric images with associated geometry by making use of a collection of posed single-view 2D images at training time. This transformation is achieved by conceptualizing the surface as the zero-level set of a neural implicit signed distance function. To achieve our high-fidelity geometric synthesis, we first introduce our augmented triplane representation. Then, we introduce our SDF-based volume rendering. Finally, we describe an SDF depth-consistency constraint, which is used to enhance SDF learning. Figure 1 displays our overall pipeline.

Augmented triplane representation. Our method augments the original EG3D triplane representation with sampling position $\boldsymbol{p}$. According to the sampling position $\boldsymbol{p}$, we retrieve the corresponding feature vector $[F_{xy}(\boldsymbol{p}), F_{xz}(\boldsymbol{p}), F_{yz}(\boldsymbol{p})]$ via bilinear interpolation. In addition, the position $\boldsymbol{p}$ is processed by a position embedder $PE(\cdot)$ that employs multi-level sine and cosine functions similar to NeRFs [26]:

$$PE(a) = [a, \gamma_0(a), \gamma_1(a), \ldots, \gamma_{L-1}(a)], \qquad (1)$$

where $\gamma_k(a) = [\sin(2^k \pi a), \cos(2^k \pi a)]$, L is a hyper-parameter that controls the maximum encoded frequency, and a represents each of the three different spatial dimensions of $\boldsymbol{p}$. $\boldsymbol{p}$ is defined as a vector since it represents the position in 3D space. Each component of $\boldsymbol{p}$ (i.e., p_x, p_y, p_z) corresponds to a different spatial dimension.

The function $\gamma_k(a)$ is a positional encoding function that takes a scalar value a and returns a 2D vector representation of the sine and cosine of $2^k \pi a$. This function is used for positional encoding to capture frequency information up to a maximum frequency defined by the hyper-parameter L.

The augmented triplane representation is formed by concatenating the triplane features $F_{xy}(\boldsymbol{p})$, $F_{xz}(\boldsymbol{p})$, and $F_{yz}(\boldsymbol{p})$ with the positional encoding $PE(p_x)$, $PE(p_y)$, and $PE(p_z)$. This augmented representation enables the model to capture high-frequency details by combining the local geometric features with positional encoding information. The

absence of the positional encoder destabilizes the training process, often resulting in model collapse (see supplementary material for results).

SDF-based volume rendering. The augmented tri-plane representation is directed to a shallow MLP to learn the SDF value $s(\boldsymbol{p})$ and RGB color $\mathbf{c}(\boldsymbol{p})$ for point $\boldsymbol{p}$. The SDF value represents the distance to the surface, providing an accurate depiction of its geometry. To convert the SDF value $s(\boldsymbol{p})$ into a density field σ, we follow VolSDF [43] and use the following Laplace transformation:

$$\sigma(s(\boldsymbol{p})) = \begin{cases} \frac{1}{2\beta} \exp\left(\frac{s(\boldsymbol{p})}{\beta}\right) & \text{if } s(\boldsymbol{p}) \leq 0 \\ \frac{1}{\beta}\left(1 - \frac{1}{2}\exp\left(-\frac{s(\boldsymbol{p})}{\beta}\right)\right) & \text{if } s(\boldsymbol{p}) > 0 \end{cases}, \tag{2}$$

where β is a parameter, which can be fixed or learned. Based on the volumetric integration, the rendered RGB color for a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is calculated as follows:

$$C(\mathbf{r}) = \sum_{i=1}^{M} T_i(1 - \exp(-\sigma_i \delta_i))\mathbf{c}_i, \tag{3}$$

where $\mathbf{o}$ is the camera position, $\mathbf{d}$ is the ray direction, $T_i = \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$ and $\delta_i = t_{i+1} - t_i$ is the distance between adjacent sampled points. For simplicity, we use σ_i and $\mathbf{c}_i$ to denote $\sigma(s(\boldsymbol{p}_i))$ and $\mathbf{c}(\boldsymbol{p}_i)$ respectively, which mean the color and density value at the *i-th* sampling point $\boldsymbol{p}_i$ along ray $\mathbf{r}$. In a similar way, we compute the rendered distance as follows:

$$d(\mathbf{r}) = \sum_{i=1}^{M} T_i(1 - \exp(-\sigma_i \delta_i))t_i. \tag{4}$$

SDF depth consistency. It has been shown in Geo-Neus [11] that there can exist a gap between the rendered image and the true surface and it is important to introduce explicit constraints to optimize the SDF network. Therefore, Geo-Neus introduces sparse points and multi-view photometric consistency to achieve this in the multi-view setting when multiple images are available for each object during training. However, these two constraints are obviously not available in our *single*-view GAN setting. To reduce the geometry bias caused by volumetric integration, the 3D point computed from the rendered distance $d(\mathbf{r})$ in Equation 4 should be located on the zero-level set of the SDF network. Thus, according to the rendered distance $d(\mathbf{r})$, its corresponding 3D point $\mathbf{p}_{d(\mathbf{r})}$ is computed as:

$$\mathbf{p}_{d(\mathbf{r})} = \mathbf{o} + d(\mathbf{r})\mathbf{d}. \tag{5}$$

Since the above 3D point should be approximately on the geometry surface, the SDF value of this point should be approximately zero. Thus, we define an SDF constraint as:

$$\mathcal{L}_s = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} |s(\mathbf{p}_{d(\mathbf{r})})|, \tag{6}$$

where $\mathcal{R}$ denotes all rays for the current camera pose. During training we aim to minimize the above loss.

3.3. Training GeoGen

The SDF-based GeoGen model uses dual discrimination during training, evaluating both the neurally rendered low-resolution 2D image and the super-resolved 2D image. The generative model takes only 2D images as input, and the discriminator encourages both the low-resolution and super-resolved synthesized 2D images to match the distribution of real images. This ensures the consistency between the super-resolved images and the neural rendering, facilitating our method to achieve high-quality high-resolution rendering results. In addition, the SDF depth consistency loss is imposed during training to promote geometric consistency. The model can then effectively learn to capture accurate geometry information from the 2D images, leading to more precise and reliable 3D reconstructions. Our overall loss is:

$$\mathcal{L} = \mathcal{L}_{dis} + \lambda \mathcal{L}_s, \tag{7}$$

where $\mathcal{L}_{dis}$ is a GAN loss computed using dual discrimination and λ is a weighting applied to the SDF constraint. Empirically we find that directly training our model from scratch is challenging. We suspect that the introduced learnable parameter β in Equation 2 prevents the StyleGAN2-based feature generator from learning effective features. In addition, the SDF constraint requires good geometry initialization, which is not possible to obtain in the early phase of training. Therefore, we design a learning strategy to train our model in which the β parameter of the Laplace density distribution is fixed to stabilize the early learning of our generative model.

Specifically, the significant part of this training process involves managing the β parameter of the Laplace transformation in Equation 2, which directly influences the learning of the SDF network. The β parameter remains fixed for the first N iterations to allow the SDF network to focus on learning coarse geometry. This enables the learning of the StyleGAN2-based generator to produce stable view synthesis. After N iterations, we make the β a learnable parameter to increase the ability of the model to capture finer-scale surface details. As previously mentioned, the SDF constraint should also be carefully managed. We achieve this by controlling the weight λ in Equation 7, where it is initially set to 0, and then increased to 0.1 after N iterations. As a result, our geometry optimization is conducted in a quasi coarse-to-fine fashion, *i.e.* N iterations, our Geo-Gen learns coarse geometry and then after this, the SDF constraint can concentrate on surface detail refinement.

4. Synthetic human head dataset

Existing methods typically train their models on high resolution human face datasets such as Flickr-Faces-HQ

Figure 2. Examples from our synthetic human dataset. We display rendered images on top and pseudo 3D ground-truth below.

(FFHQ) [19]. However, FFHQ only contains a limited range of captured viewpoints (*i.e*. no backs of heads) and has no 3D ground-truth, hence the need for our synthetic dataset. There are other synthetic datasets, such as ShapeNet Cars [6], which have ground-truth 3D meshes but are not realistic looking.

To address this, we created a new dataset of semi-realistic synthetic human heads which is generated based on the work of Wood et al. [41]. Our dataset features images of different synthetic individuals with diverse facial features, body morphologies, clothing, and hair styles. Crucially, unlike FFHQ which primarily captures frontal views, our dataset includes images across the full azimuth range, ensuring comprehensive representation of heads from all sides. This approach not only fills a critical gap in available resources but also shifts the focus towards the quality of the mesh, a vital aspect for advancing the field of 3D generative modeling.

For our dataset, we randomly generate 10 images of 512×512 for each of 19,800 identities, ensuring a comprehensive set of different views, encompassing full azimuthal coverage and utilize multi-view stereo and surface reconstruction techniques to establish pseudo ground-truth meshes. To generate a pseudo ground-truth mesh for quantitative evaluation of 3D reconstruction metrics we use the ACMP multi-view stereo approach from [42] and Poisson surface reconstruction [22] to reconstruct the full head geometry. Example images can be found in Figure 2. A subset of images from our synthetic dataset will be made available upon acceptance.

5. Experiments

Here we present qualitative and quantitative results comparing GeoGen to existing methods. For the baseline EG3D model, we retrained it on each of the evaluation datasets so that the training settings were consistent with our approach (*e.g*. the same number of training epochs). Implementation details are provided in the supplementary material.

Dataset	Method	FID↓	KID↓	ID↑
FFHQ	GRAF	79.20	55.00	-
	PiGAN	83.00	85.80	0.67
	GIRAFFE	31.20	20.10	0.64
	HoloGAN	90.90	75.50	-
	StyleSDF	11.50	2.65	-
	EG3D	4.86	0.0053	0.77
	EG3D (rebalanced)	**4.70**	**0.0044**	**0.79**
	EG3D**	5.70	0.0054	0.76
	GeoGen	5.40	0.0049	0.75
Synthetic Heads	EG3D**	5.90	0.65	**0.69**
	GeoGen	**5.10**	**0.0038**	**0.69**
ShapeNet Cars	GIRAFFE	27.30	1.70	-
	Pi-GAN	17.30	0.93	-
	EG3D	2.75	0.0054	-
	EG3D**	2.90	0.0043	-
	GeoGen	**2.50**	**0.0028**	-

Table 1. Comparative analysis of different generative models on FFHQ, our Synthetic Heads, and ShapeNet Cars datasets using standard 2D metrics. Our model surpasses EG3D [5] and other leading models in both FID and ID metrics for the Synthetic Heads and ShapeNet V1 datasets. However, it does not outperform EG3D on the FFHQ dataset, attributed to a lower number of training iterations due to limited computational resources. Additionally, the original number of training epochs for achieving the reported FID results in EG3D is not specified by its authors. GeoGen was not included in training on the FFHQ rebalanced dataset due to its unavailability during the training period. ** indicates our retraining with far fewer iterations and computation power.

5.1. Datasets

We perform experiments on Flickr-Faces-HQ (FFHQ) [19], ShapeNet Cars [6], and our synthetic human dataset described previously. Each provide distinct, valuable resources for training and evaluating 3D-aware generative models. The FFHQ dataset consists of high-quality real 2D face images. It contains over 70,000 1024×1024 resolution images. ShapeNet Cars provides images for a variety of car models imaged from different viewpoints. The dataset we used for training contains 2,100 different car instances, each with 20 images from different viewpoints.

5.2. Quantitative results

We adopt the widely used Frechet Inception Distance (FID) [17] and Kernel Inception Distance (KID) [2] metrics to measure the image synthesis quality of our GeoGen approach. We also assess multi-view facial identity consistency (ID) by calculating the mean Arcface [8] cosine similarity score between pairs of views of the same synthesized face rendered from random camera poses. We report the results of our retrained EG3D baseline using the same training conditions and our GeoGen model on the three different datasets in Table 1. Our improved results show that our GeoGen can achieve better image synthesis results on

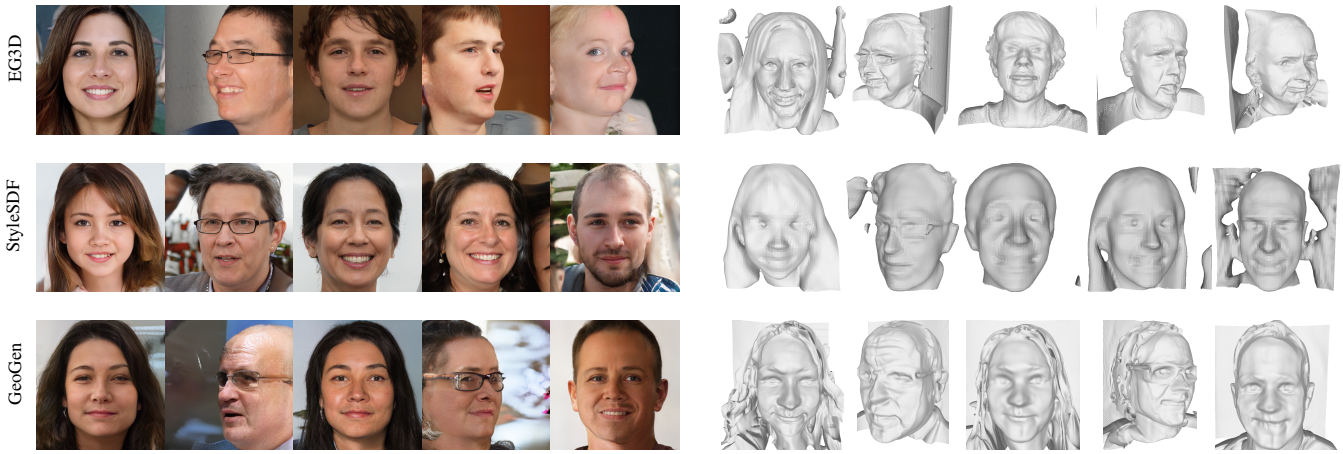

Figure 3. Sampled images and meshes from EG3D, Style SDF, and our GeoGen approach on FFHQ. GeoGen meshes display smoothness, anatomical accuracy, and detailed facial features. In contrast to EG3D and Style SDF, GeoGen synthesizes finer geometric detail.

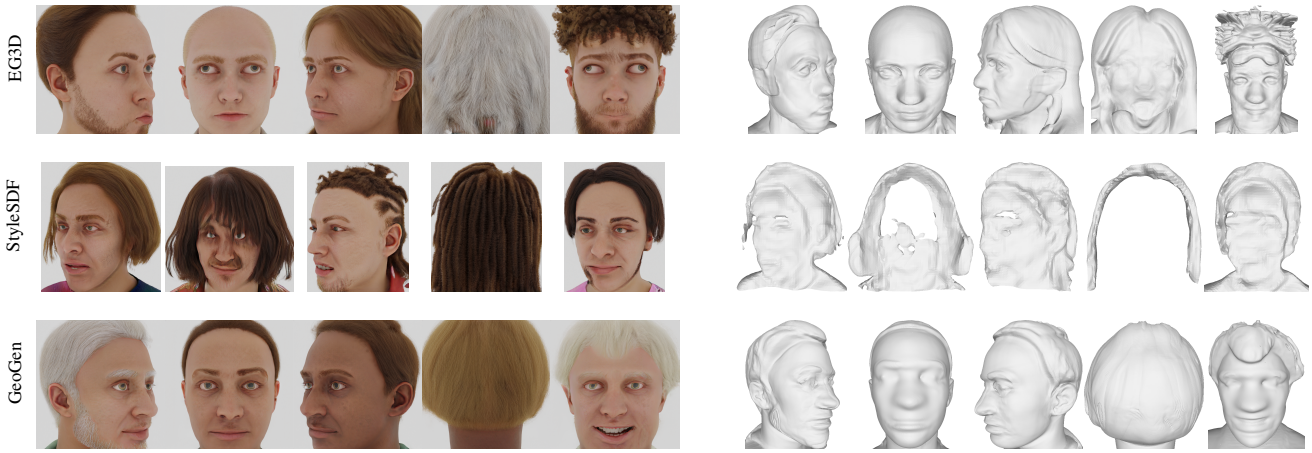

Figure 4. Sampled images and meshes from EG3D, StyleSDF, and our GeoGen approach trained on our synthetic human head dataset. GeoGen results in fewer overt visual artefacts and more faithfully captures the backs of objects (*e.g.* see second last column). While the 2D images from the competing methods look plausible, the underlying 3D mesh is not always consistent.

synthetic humans and ShapeNet Cars datasets.

An important feature of our approach is its ability to generate accurate meshes from a single image. However, it is difficult to evaluate the *geometric* quality of generative models on real images as ground-truth 3D shape information is challenging to obtain. Instead, it is possible to obtain the ground-truth meshes for both synthetic datasets that we use. To evaluate the generated meshes of different methods quantitatively, we leverage the GAN inversion technique PTI [33]. Then, given an image from the test set dataset, we can estimate the corresponding latent code by PTI. With the latent code, we can generate both the synthesized image and mesh. In this way, we can compute a range of 3D evaluation metrics that compare the differences between the synthesized mesh and ground-truth mesh to measure the geometry fidelity. Results are presented in Table 2, where we observe that our GeoGen outperforms EG3D.

5.3. Qualitative results

Here we present qualitative results where we compare GeoGen to existing methods. In Figures 5 and 7 we compare 2D image synthesis of different methods via GAN inversion. We observe that GeoGen results in outputs that more closely match the input image. In Figure 7 we observe that GeoGen captures details such as the spacing between the car body and wheel and, in some instances, even the handles on the doors of the cars. Finally, in Figures 3 and 4 we display sampled outputs (*i.e.* not inversions).

6. Discussion

Our evaluation shows the competitive performance of our proposed GeoGen model, both qualitatively and quantitatively. To gain deeper insight into the effectiveness of our approach, we employed a suite of metrics that assess both the 2D and 3D aspects of the images and meshes gener-

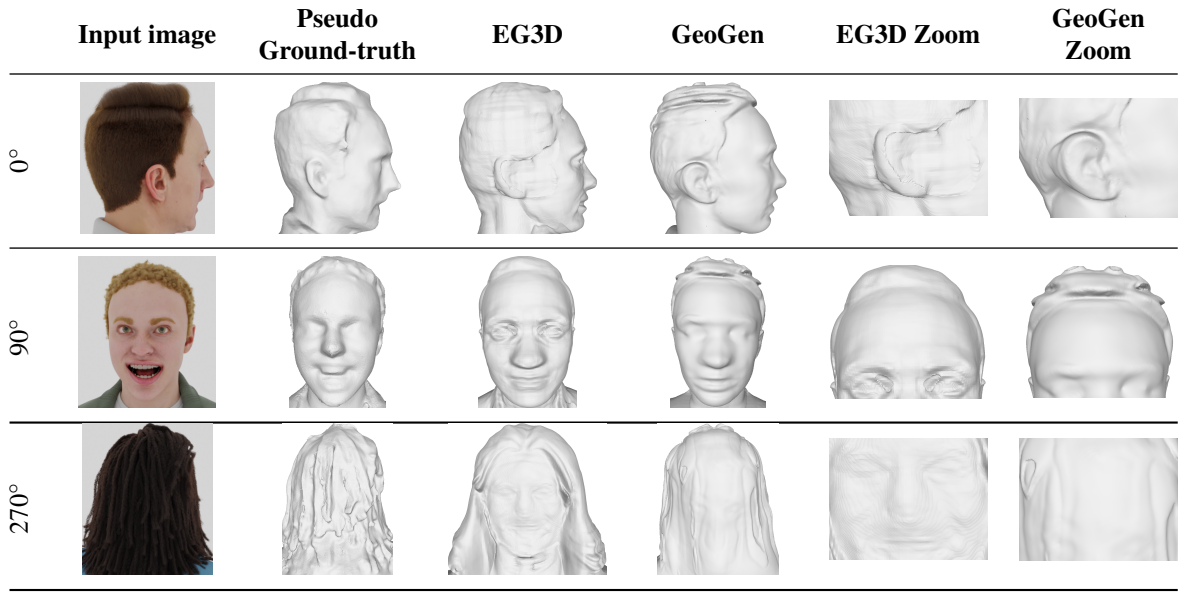

Figure 5. Inversion Results for EG3D and GeoGen Models: The figure presents a comparison at 0°, 90°, and 270° angles to highlight variations in the reconstruction of facial features by the two models.

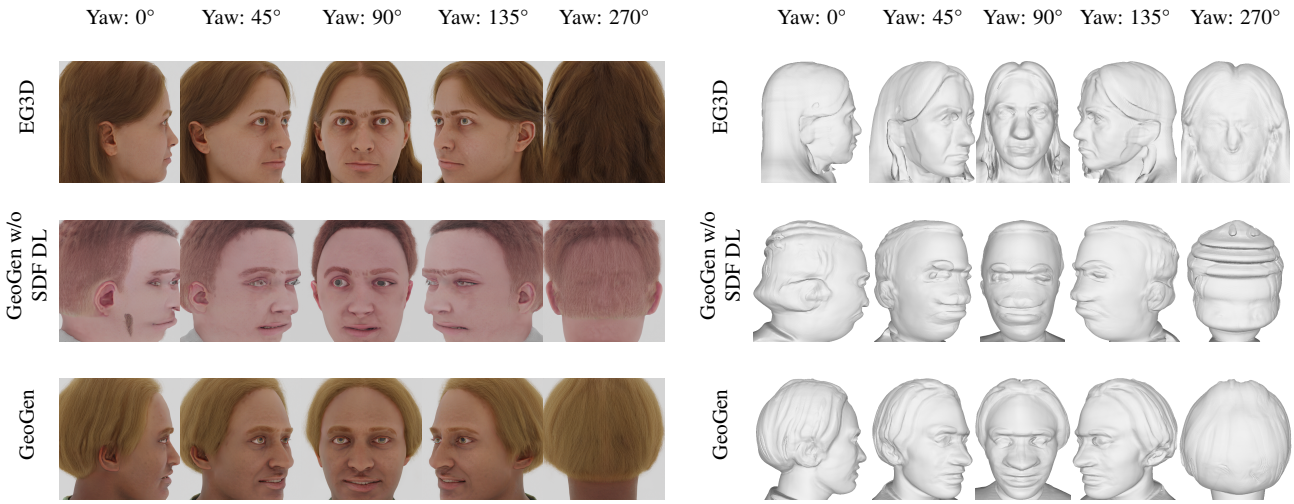

Figure 6. Comparison of EG3D and GeoGen, with and without SDF Depth Loss (SDF DL) constraints, showing sampled images from models trained on our synthetic human images. These examples highlight GeoGen's ability to represent finer geometric details, *e.g.* the ears have more detail than those generated by EG3D. We also observe a failure for EG3D in the top right, where the back of the head contains facial geometry. More qualitative results highlighting the differences in the use of the SDF depth loss are shown in the supplementary.

ated by our model. Two quantitative performance areas are of particular note: the synthesis of high-quality 2D images and precise 3D geometric predictions. Our model competes closely with EG3D [26] in terms of 2D metrics, outperforming both StyleSDF [32] and GRAF [34]. This demonstrates our model's ability to generate high-fidelity 2D images.

Table 2 showcases a systematic comparison between GeoGen and EG3D, revealing the advantages of incorporating Signed Distance Functions (SDF) and SDF depth constraints during training. The lower Chamfer Distance for GeoGen compared to EG3D for both Cars and synthetic human heads is indicative of a more precise alignment between the reconstructed points and corresponding points in the ground-truth. This highlights an improved precision in point-to-point correspondence which is an essential part of 3D reconstruction. The Earth Mover's Distance, another vital metric in understanding the geometrical congruence between shapes, is also consistently lower for GeoGen. This indicates that the shapes are more similar, requiring fewer alterations to match the ground-truth, thus showing an underlying efficiency in GeoGen's modeling approach. Finally, the Mean Surface Distance adds to the evidence of

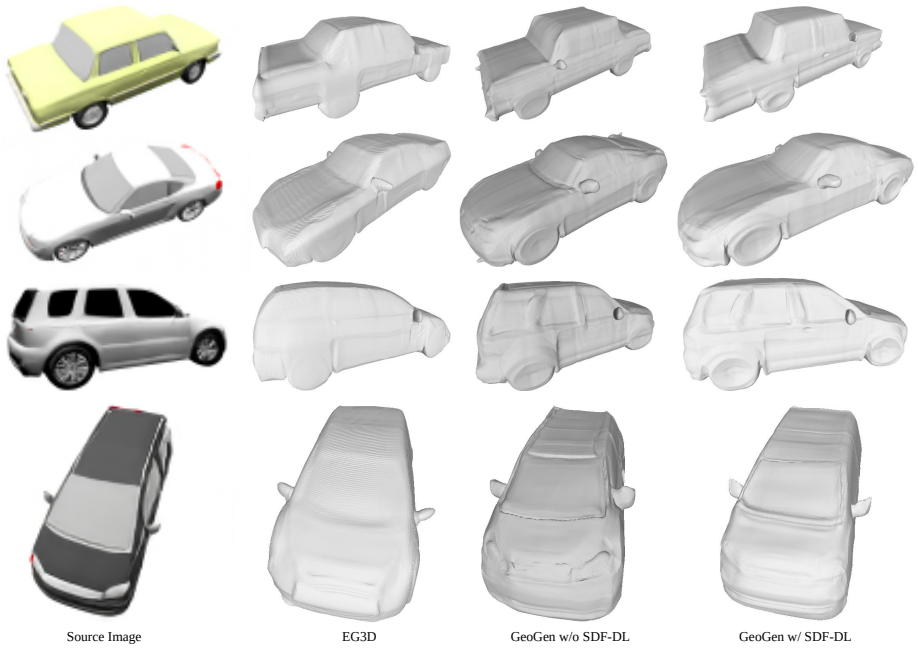

Figure 7. Comparison of mesh predictions on ShapeNet Cars. Meshes are obtained by inverting the source image to derive latent codes. EG3D meshes display diminished shape fidelity and surface detail. Using SDF constraints in GeoGen improves detail, evident around car wheels and windows. Results for GeoGen without SDF constraints are also shown for context.

ShapeNet Cars					
Method	Chamfer↓	MSE↓	HD↓	EMD↓	MSD↓
EG3D	0.31	0.31	0.85	0.44	0.33
GeoGen w/o SDF&Depth Loss	0.27	0.28	**0.77**	0.42	0.31
GeoGen	**0.25**	**0.27**	**0.77**	**0.40**	**0.29**
Synthetic Heads					
Method	Chamfer↓	MSE↓	HD↓	EMD↓	MSD↓
EG3D	0.21	0.29	0.65	0.54	0.35
GeoGen w/o SDF& Depth Loss	0.19	0.29	0.59	0.45	0.26
GeoGen	**0.17**	**0.27**	**0.56**	**0.43**	**0.24**

Table 2. Comparison of different 3D reconstruction metrics for generative models on *ShapeNet Cars* and our *Synthetic Heads* dataset. We report averages for MSE, HD, and MSD metrics. Variations of GeoGen without the SDF and Depth Loss constraints are also shown. Best methods for each dataset are bolded.

GeoGen's superiority, as it also yields consistently lower values. The implication here is a closer similarity between the reconstructed and target shapes, providing further evidence for GeoGen's effectiveness.

The utilization of the SDF in GeoGen ensures better geometric consistency in the reconstruction, as it leverages the implicit representation of the mesh's surface. GeoGen, with its additional depth constraints, preserves topology and fine details that are often overlooked with conventional generative techniques like EG3D (see Figure 6). It is also noteworthy that these numerical advantages, though significant, do not fully represent the perceptual quality of the reconstructed models. Qualitative evaluations indicate that models generated by GeoGen often appear more realistic and accurate, underscoring GeoGen's advantage in bridging quantitative performance with perceptual realism.

Limitations. Our GAN-based approach, like others, requires posed images for training. Camera poses can be estimated similar to methods used in FFHQ. While we aim to align the expected depth with the SDF's zero-level set, extending the SDF consistency loss to other points along the ray could theoretically enhance geometric accuracy. However, this would substantially increase computational load. There are also inherent limitations in learning-based methods, such as potential bias from unrepresentative training data, notably in web-scraped human face images.

7. Conclusion

We presented GeoGen, a novel 3D-aware generative model for synthesizing high-quality 2D images with associated accurate 3D geometry, that is trained from 2D images. GeoGen outperforms established methods on several performance metrics. By harnessing the power of neural implicit representations and neural signed distance functions, we have developed a solution that delivers both quality and versatility in the context of 3D representation learning. In addition, we presented a new synthetic human head dataset for training and quantitatively evaluating 3D generative models. GeoGen moves us closer to the goal of enriching fields such as character animation, gaming, and virtual reality with plausible 3D geometry from single input images. Our results affirm the potential of our approach and its relevance in this rapidly evolving field.

References

[1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360°. In *Conference on Computer Vision and Pattern Recognition*, 2023. 2, 15

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 5

[3] Eric Chan. Efficient geometry aware 3d network. *Computer Vision and Pattern Recognition*, 2022. 11, 12, 13, 16, 17, 18

[4] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 5

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv:1512.03012*, 2015. 5, 14

[7] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Conference on Computer Vision and Pattern Recognition*, 2019. 5

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

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 2021. 2

[11] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 2022. 1, 2, 4, 11

[12] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *Conference on 3D Vision*, 2017. 2

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

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 1, 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *International Conference on Computer Vision*, 2015. 2

[16] Paul Henderson, Vagia Tsiminaki, and Christoph H Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *Conference on Computer Vision and Pattern Recognition*, 2020. 2

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 2017. 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020. 2

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1, 5, 13, 15

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Conference on Computer Vision and Pattern Recognition*, 2020. 3

[21] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 2021. 1

[22] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Fourth Eurographics symposium on Geometry processing*, 2006. 5

[23] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Eurographics symposium on Geometry processing*, 2006. 11, 13

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. 2

[25] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1

[26] Ben Mildenhall, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Ricardo Martin-Brualla. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020. 1, 2, 3, 7

[27] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *International Conference on Computer Vision*, 2019. 2

[28] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in Neural Information Processing Systems*, 2020.

[29] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Conference on Computer Vision and Pattern Recognition*, 2021. 2

[30] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *International Conference on Computer Vision*, 2021. 1

[31] 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 *Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 11

[32] Jeong Joon Park, Philip Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 7

[33] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *Transactions on Graphics*, 2022. 2, 6

[34] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 2020. 2, 7

[35] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *Advances in Neural Information Processing Systems*, 2022. 2

[36] Minjung Shin, Yunji Seo, Jeongmin Bae, Young Sun Choi, Hyunsu Kim, Hyeran Byun, and Youngjung Uh. Ballgan: 3d-aware image synthesis with a spherical background. In *International Conference on Computer Vision*, 2023. 1

[37] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2

[38] E. Tov, J. Doe, and A. Smith. Pivotal tuning inversion. *Computer Graphics Forum*, 2023. 15

[39] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems*, 2021. 1, 2

[40] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Conference on Computer Vision and Pattern Recognition*, 2023. 2, 15

[41] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *International Conference on Computer Vision*, 2021. 5

[42] Qingshan Xu and Wenbing Tao. Planar prior assisted patchmatch multi-view stereo. In *AAAI Conference on Artificial Intelligence*, 2020. 5, 11, 13

[43] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 2021. 1, 2, 3, 4

Appendix

The foundation of our model relies on the official implementation of Enhanced Generative 3D Models (EG3D) [3]. We utilized R1 regularization, assigning a gamma = 1 for the synthetic humans and FFHQ dataset based on the input image size of 512 x 512 and batch size of 32 across 8 v100 GPUs, following the same hyperparameter tuning of EG3D. For ShapeNet Cars, we adopted a gamma value of 0.3 based on the 128 x 128 resolution and batch size of 32 [11]. Our model employs the same architecture as StyleGAN2 [31], composed of a mapping network with 8 hidden layers, and output convolutions yielding 96 feature maps. Following the EG3D protocol, these are then reshaped into 3 planes of 256 x 256 x 32 [3].

.1. GeoGen training

During the initial training of GeoGen for the FFHQ and Synthetics dataset, the model was trained end-to-end, a process that necessitated unique handling of the SDF depth consistency loss. For the first 10,000 epochs, we set the beta value for the Laplace density distribution to 0.1 and refrained from making it learnable, as our end-to-end model would not have been able to learn the best beta value at this stage [11]. This approach allowed the model to first learn the optimal geometry and SDF depth map. In contrast, StyleSDF had to introduce a two-stage training process precisely because their pipeline was not trained end-to-end. They consistently used a learnable beta parameter for the Laplace density distribution throughout their training, as their method required more flexibility in the control of the SDF consistency loss.

The Laplace beta value plays a crucial role in the SDF network as it controls the shape of the Laplace distribution, influencing how the model penalizes deviations from the expected SDF values. A lower beta value produces a wider distribution, allowing for a larger spread of SDF values, and a higher beta value tightens the distribution, constraining the SDF values more strictly. This ability to control the distribution of SDF values enables fine-tuning of the model's sensitivity to inconsistencies in the SDF depth, a key aspect of the learning process. After the generator in our model showed improvement in rendering, depth maps, and underlying geometry, we activated the SDF constraint for depth map regularization and introduced the learnable beta parameter for the remaining 10,000 epochs. This allowed us to dynamically adapt the SDF consistency loss and fine-tune the model's learning of SDF depth.

Both EG3D and GeoGen models underwent training for 20,000 epochs for the FFHQ and Synthetics data, while for the ShapeNet dataset, training was conducted for 10,000 epochs. The batch size for all models was 18, with the discriminator's learning rate at 0.002 and the generator's at 0.0025. The training was carried out using 4 NVIDIA P100, while an RTX 2080 and RTX 4090 were used for inference during inversions and sample generation. Our end-to-end training approach, including the specific handling of the Laplace beta value, was central to our method's effectiveness in learning SDF depth. It allowed us to combine the flexibility needed in the early stages of learning with the precision required in later stages, reflecting a sophisticated understanding of the role that SDF plays in the generative process.

.2. SDF and color network and surface rendering

The resulting embedding from the augmented spatial representation is fed into the SDF (Signed Distance Function) network. This network utilizes the embedded position to query the SDF value at a specific point, which gives precise information regarding the distance to the nearest surface within the 3D space. The understanding of these distances is crucial in the reconstruction of 3D objects, as it provides detailed insights into the geometry and the underlying complexities of the data being modeled.

Once the SDF network receives and processes the embedded position, the computed SDF values are further handled by the color network. This auxiliary network takes the SDF values and translates them into the corresponding color values for the rendered 3D object. The direct utilization of SDF values as input for the color network establishes a coherent link between the geometric structure and visual appearance of the object. Both the SDF and color networks are built with a single hidden layer comprising 64 hidden units and leverage a soft plus activation function. This structure ensures smooth transitions and optimal gradient flow within the networks. For the transformation of the SDF into tangible density, a specific surface rendering technique has been applied. The sampling strategy is carefully chosen and tailored to different datasets, such as using 48 uniformly spaced samples and 48 importance samples per ray for the FFHQ dataset, and 64 of each for ShapeNet cars and Synthetics data.

In combination, these elements forge an intricate pipeline that integrates spatial features and coordinates, through a positional encoder, with the SDF and color networks. The methodology's architecture ensures a nuanced and true-to-life representation across a multitude of datasets. The implementation of a positional encoder has further enhanced the SDF network's capacity to grasp and replicate complex 3D geometries. The employment of SDF networks for surface rendering has led to a more sophisticated and resilient interpretation of various datasets.

.3. Reconstruction of pseudo ground truth meshes

To reconstruct pseudo ground truth meshes we use Planar Prior Assisted PatchMatch Multi-View Stereo (ACMP) [42] and Poisson surface reconstruction [23]. Example pseudo

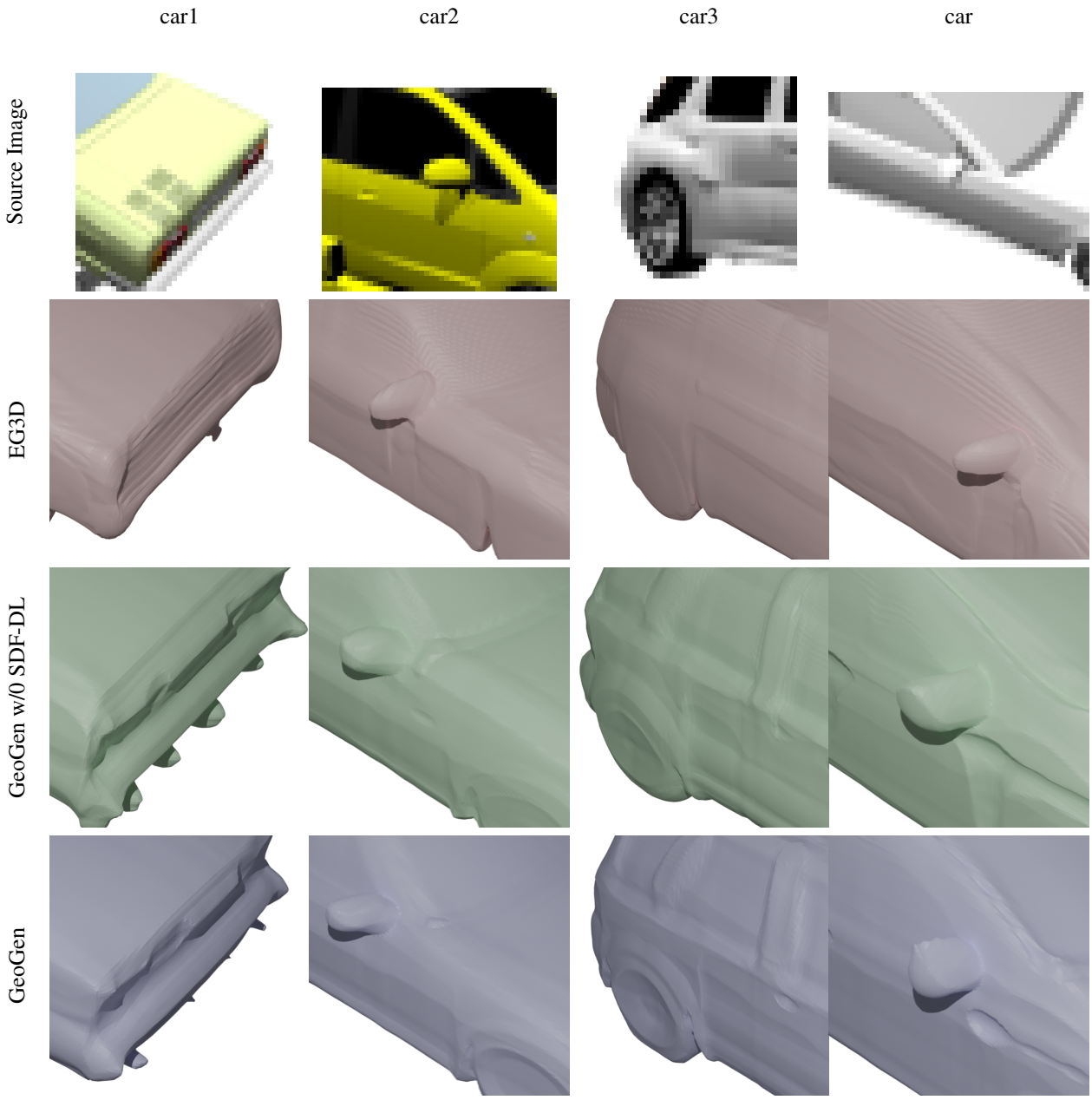

Figure A1. A detailed comparison between EG3D and GeoGen in the context of ShapeNet cars inversion of meshes, emphasizing the differences in the geometric representation and rendering capabilities of both methods. The samples underscore the advanced efficacy of GeoGen in capturing and reconstructing intricate geometric details within the car models, even at granular levels. This superiority is attributed to the integration of the Signed Distance Function (SDF) network along with the SDF depth consistency loss within GeoGen's architecture. The SDF approach provides a continuous and differentiable representation of the car's surface, enabling more precise and robust alignment with the observed data. This contributes to better capturing of fine geometrical nuances and results in more accurate reconstructions. Conversely, the EG3D [3] method's rendered meshes reveal a deficiency in portraying granular details, leading to a more approximate and less nuanced depiction of the vehicles.

ground truth meshes are shown in Figure A5. Recognizing the challenge of depth estimation in low-textured areas, which typically exhibit strong planarity, ACMP makes use of planar models in conjunction with the PatchMatch algorithm. By embedding planar models into PatchMatch MVS via a probabilistic graphical model, our approach introduces a multi-view aggregated matching cost. This novel cost function takes both photometric consistency and pla-

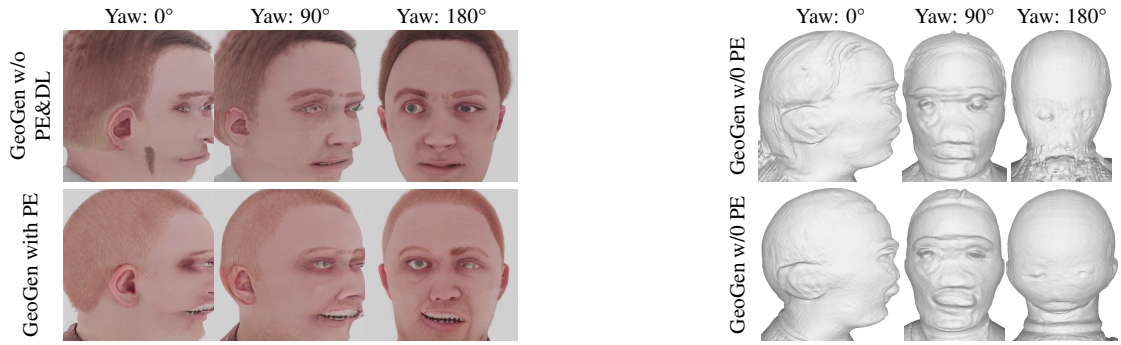

Figure A2. This caption accompanies a series of synthetic images generated by the GeoGen model operating without a positional encoder. The figures on the left illustrate the model's output at different yaw angles, showcasing its ability to render facial features from various perspectives. On the right, the corresponding mesh structures are displayed, providing a deeper insight into the model's geometric rendering capabilities. These results were captured prior to the point of model collapse, highlighting the model's performance and limitations in the absence of positional encoding. This comparison not only demonstrates the visual output of the model but also underscores the critical role of positional encoding in maintaining structural integrity and realism in the generated images and meshes.

nar compatibility into consideration [42], thus accommodating both non-planar and planar regions. This method has demonstrated its capability to recover depth information in areas of extremely low texture, efficiently leading to high completeness in 3D models.

The problem of surface reconstruction from oriented points is cast as a spatial Poisson problem using Poisson surface reconstruction. This formulation's advantage is its simultaneous consideration of all points without the need for heuristic spatial partitioning or blending, which enhances resilience to data noise [23]. The use of a hierarchy of locally supported basis functions and the reduction of the solution to a well-conditioned sparse linear system makes this approach computationally efficient.

By seamlessly integrating ACMP with Poisson surface reconstruction, we've crafted a novel method for 3D model reconstruction. The fusion of these techniques allows us to address the complexities and subtleties of 3D modeling, particularly in challenging scenarios where noise and low texture might otherwise impede reconstruction. The reconstructed pseudo-ground truth meshes generated by this combined approach are a testament to its effectiveness, signifying an exciting advancement in the realm of 3D modeling and a promising avenue for further exploration and optimization.

.4. Results without positional encoder

Here we explore causes behind the collapse of the GeoGen model, specifically when trained without the aid of positional encoding in the context of Neural Radiance Fields (NeRF) and GAN training. The absence of positional encoding can lead to several critical issues (see Figure A2). Firstly, in GAN training, the phenomenon of mode collapse becomes more pronounced. This is where the generator starts producing a limited variety of outputs, failing to capture the complex data distribution. Secondly, the intrinsic characteristics of NeRF, which rely heavily on precise spatial information to render 3D scenes accurately, are compromised without positional encoding. This results in the model's inability to effectively learn and represent high-frequency details, leading to a loss of detail and realism in the generated images. Lastly, positional encoding plays a vital role in stabilizing the training process by providing a more detailed and nuanced understanding of spatial relationships in the data. Its absence can result in unstable training dynamics, ultimately causing the model to collapse, particularly evident in our observations post epoch 11000. This highlights the essential nature of positional encoding in maintaining the stability and efficacy of models like GeoGen, especially in complex applications involving synthetic human images and 3D rendering.

A. Datasets

A.1. FFHQ and rebalanced FFHQ

Our modeling framework originally utilized the "in-the-wild" version of the FFHQ dataset, a comprehensive collection of uncropped, original PNG human images sourced from Flickr, as documented by Karras et al. (2019) [19]. To adapt these images for our purposes, we employed a sophisticated face detection and pose-extraction system [3], allowing us to determine the face area and label each image with its corresponding pose. The images were then cropped to approximate the dimensions of the original FFHQ dataset. We assumed fixed camera intrinsics for all images, with a focal length 4.26 times the image width, mimicking a standard portrait lens [3]. After removing a small number of images where face detection proved unsuccessful, our final dataset comprised 69,957 images.

In our reporting, we include the 2D performance metrics of models trained on the Rebalanced FFHQ dataset, particularly focusing on the outcomes from NVIDIA-trained mod-

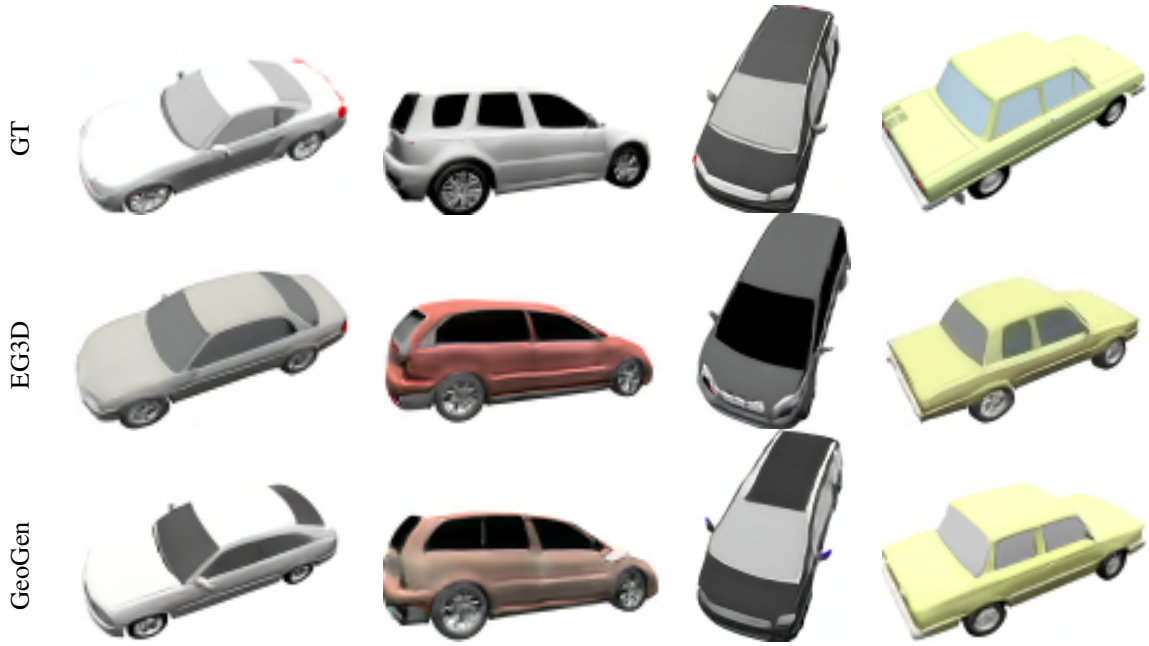

Figure A3. Comparison of EG3D and GeoGen inversion results using held-out images from the ShapeNet Car test set. GeoGen results more closely resemble the input ground truth image (GT).

els. The Rebalanced FFHQ dataset, known for its broader diversity in facial orientations, plays a crucial role in enhancing the model's capability to understand and replicate human facial features from various angles. This dataset is especially valuable for models that need to handle a wide range of facial geometries, such as those used in advanced image generation and recognition tasks.

While we present these metrics to showcase the performance improvements facilitated by the Rebalanced FFHQ dataset, it's important to note a limitation in the available data. NVIDIA, the entity responsible for training these models, has not provided detailed information regarding the number of epochs, specific training methodologies, or other intricate details of the training process. This lack of detailed training information could potentially impact the reproducibility and further optimization of these models.

Understanding the training duration (measured in epochs) and the specific methodologies employed is crucial for comprehensively evaluating a model's performance and for making informed comparisons with other models. The absence of this information leaves a gap in fully understanding how the Rebalanced FFHQ dataset impacts model performance compared to the original FFHQ dataset. Despite this, the reported 2D metrics still offer valuable insights into the enhanced capabilities of models trained on the Rebalanced FFHQ dataset, highlighting their improved proficiency in handling diverse facial features and orientations.

A.2. ShapeNet V1

We utilized the ShapeNet V1 Cars dataset for additional validation, rigorously comparing methodologies on a specific subset that includes 128 renderings of synthetic cars [6]. This carefully curated dataset offers a robust platform for assessing performance across various viewing angles, enabling a comprehensive evaluation of 3D reconstruction and rendering techniques.

The ShapeNet dataset, as employed in our setup, builds on prior research and consists of 2,100 car images captured from 50 different perspectives [6]. The multi-angle images provide an ideal scenario to analyze geometric consistency, shadow rendering, and surface texturing. Similar to the preprocessing applied to the FFHQ dataset, our approach to the ShapeNet data followed established protocols, maintaining the integrity and original characteristics of the images. Unlike other methodologies that might use augmentation or mirror images, we consciously chose not to apply these techniques to preserve the authenticity of the data and ensure a more accurate assessment of the models' performance [6].

A.3. Synthetic humans

Our training model also harnessed our proprietary synthetic human dataset. This extensive collection encompasses 200,000 images, representing 20,000 unique identities. Each of these identities is portrayed from only 10 viewpoints, a stark contrast to the Rodin model where each

identity was rendered from 300 diverse viewpoints [40]. Despite the significant reduction in viewpoints per identity in our dataset, our model produces high quality outputs in terms of geometry and rendering [1]. Our training approach proves that strong performance can be achieved with a more limited number of viewpoints.

A.4. Pivotal tuning inversion

In the context of our work with Pivotal Tuning Inversion (PTI), a specialized process to invert generative models like StyleGAN, we adopt a meticulous procedure to enhance the accuracy and efficiency of the inversion.

Initially, we utilize an off-the-shelf face detection solution to accurately locate and extract face regions within the test images. This process allows for precise alignment and ensures that the features of interest are adequately centered and scaled. The extracted regions are then cropped and resized to a consistent resolution of 512x512 pixels, facilitating uniform processing and analysis across different images.

Following this preprocessing stage, we implement the PTI methodology as delineated by Tov et al. [38]. This approach consists of two main stages:

1. **Fine-tuning of generator weights.** Subsequent to the initial latent code optimization, we proceed with an additional 500 iterations dedicated to fine-tuning the generator's weights. This phase is pivotal in refining the subtle details and enhancing the realism of the generated images. By adjusting the generator's parameters, we align the synthetic outputs more closely with the underlying distribution of the real data, improving both the fidelity and the perceptual quality of the inversions.
2. **Latent code optimization.** For the first 500 iterations, we focus on the optimization of the latent code, a compact representation within the model's latent space that encodes the essential features of the target image. Utilizing gradient-based optimization techniques, we iteratively refine the latent code to minimize the discrepancy between the generated image and the target. This stage ensures that the inverted model captures the essential characteristics of the face.

The combination of these two stages offers a robust and precise inversion process, enabling us to generate high-quality, detailed images that faithfully represent the original inputs. The PTI methodology, by explicitly separating the optimization of the latent code and the fine-tuning of the generator, provides a nuanced control over the inversion process, yielding superior results in terms of both accuracy and visual appeal.

A.5. Justifying the limitations in GAN inversion

In the field of Generative Adversarial Networks (GANs), particularly with advanced models like EG3D, the accuracy of GAN inversion can be inconsistent. This inconsistency can be attributed to several factors, encompassing both the inherent characteristics of the generative model and the methodologies used in the inversion process.

Firstly, the architecture and complexity of the GAN model play a crucial role. A model with limitations in its design may not capture a broad range of features effectively, leading to challenges in accurately reproducing certain types of images during inversion. For example, if the model's architecture does not account for a wide variety of facial orientations, it may struggle with accurately inverting images that fall outside of its trained norm.

Additionally, the scope and diversity of the training data are critical. A model trained on a dataset with limited variety, such as one primarily consisting of front-facing images, may not perform well in inverting images with diverse or unusual orientations. The quality and diversity of the training data directly influence the model's ability to handle a wide range of inversion tasks.

Furthermore, the model's resolution and detail capabilities are also significant. Models that generate lower-resolution images or lack fine detail might fail to accurately capture nuances in the inversion process, resulting in less precise or realistic inversions.

On the side of inversion methodologies, the efficiency of the algorithm and its approach to navigating and manipulating the latent space of the GAN are key factors. The choice of loss functions and regularization techniques within the inversion method can greatly affect the match quality between the inverted image and the original. Computational constraints can also limit the effectiveness of more resource-intensive, yet potentially more accurate, inversion methods.

In summary, the limitations in GAN inversion accuracy can be attributed to a complex interplay of factors related to both the generative model's characteristics and the inversion techniques used. Understanding and addressing these factors is crucial for improving the accuracy and reliability of GAN inversions.

A.6. Evaluation metrics

Evaluating the quality and performance of generated images is paramount in understanding the effectiveness of generative models. To this end, we employed the Fréchet Inception Distance (FID) and Kernel Inception Distance (KID), calculating these metrics for 50,000 generated images against all training images for both FFHQ and synthetic humans datasets. The calculations were performed using the implementation provided in the StyleGAN2 codebase [19], ensuring consistency with commonly accepted standards.

Our GeoGen model's KID scores were found to be 100 times lower than those of comparative models, an unexpected result that warrants careful consideration. One pos-

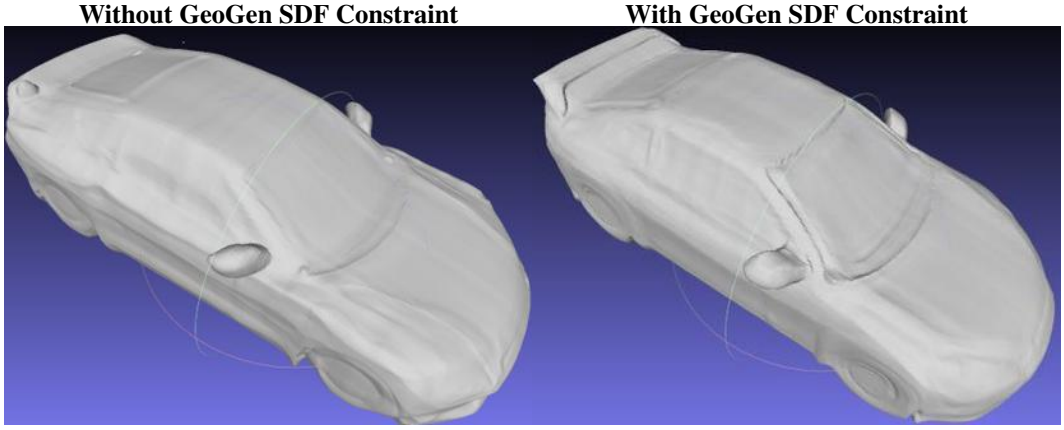

Figure A4. Comparison of models without (left) and with (right) our GeoGen SDF constraint.

sible hypothesis for this abnormality might be an alignment of specific features or particularities in the convergence behavior during the training of our model. It could also be related to the choice of hyperparameters or the data preprocessing steps that were unique to our experiment. However, these hypotheses are subject to further investigation, and the exact reason behind the unusually low KID score remains an intriguing question for future research.

Alongside the 2D image quality evaluation, we also assessed 3D geometry comparisons, adopting the Efficient Geometry Aware 3D Network (EG3D) [3] for evaluation. Our GeoGen model showed promising results relative to the EG3D model, as indicated by these metrics, both in terms of 2D image quality and 3D Chamfer distance metrics. The overall evaluation paints a comprehensive picture of our model's capabilities, but the abnormally low KID score serves as a reminder that there may always be underlying complexities and subtleties that require further exploration and understanding.

A.7. 3D reconstruction metrics

The assessment of 3D geometry is a critical aspect of our evaluation, as it reflects the ability of the generative models to faithfully reconstruct and represent the intricate geometric details of the subjects. Table 2 from the paper presents a comprehensive comparison of different 3D reconstruction metrics for generative models on ShapeNet *Cars* and Synthetic Human *Heads*. The selected metrics include Overall Chamfer Distance, Mean Squared Error (MSE), Hausdorff Distance (HD), Earth Mover's Distance, and Mean Surface Distance (MSD).

These metrics were chosen for their ability to capture various aspects of geometric fidelity. Chamfer Distance provides a measure of dissimilarity between two point sets, emphasizing both the precision and recall of the reconstructed surfaces. MSE offers insights into the mean differences between corresponding points, focusing on local accuracy. HD measures the maximum distance from a point in one set to the nearest point in the other set, highlighting global discrepancies. Earth Mover's Distance quantifies the minimum amount of work to transform one point set into the other, capturing overall distribution alignment. Lastly, MSD focuses on the mean distance between surfaces, reflecting surface smoothness and consistency.

In the process of evaluating these metrics, we scaled the generated and ground-truth meshes to fit within a unit sphere to ensure a consistent basis for comparison. We then randomly sampled 20,000 points from the meshes, repeating this process 20 times, in order to compute the mean and standard deviation of the metrics. This methodology allowed us to capture a comprehensive and statistically robust representation of the geometric quality, eliminating potential biases related to specific sampling patterns or scaling discrepancies.

The results, as shown in Table 2 of the main paper indicate that GeoGen demonstrates superior results, reflecting its ability to represent finer geometric details. The table also includes comparisons with GeoGen without SDF and DL constraints, allowing for an understanding of how specific components and constraints influence model performance. The best-performing methods for each dataset are highlighted in bold, striking a balance between quantitative performance and perceptual realism. The rigorous evaluation of these 3D metrics underscores the effectiveness of our approach and contributes to a nuanced understanding of generative modeling for complex geometric structures.

B. Additional qualitative results

In Figure A5 we present a comparison of synthetic human avatar meshes across EG3D [3] and GeoGen. It is qualitatively evident that our model, leveraging the capabilities of the Signed Distance Function (SDF) network with SDF depth consistency loss, surpasses both EG3D and StyleSDF

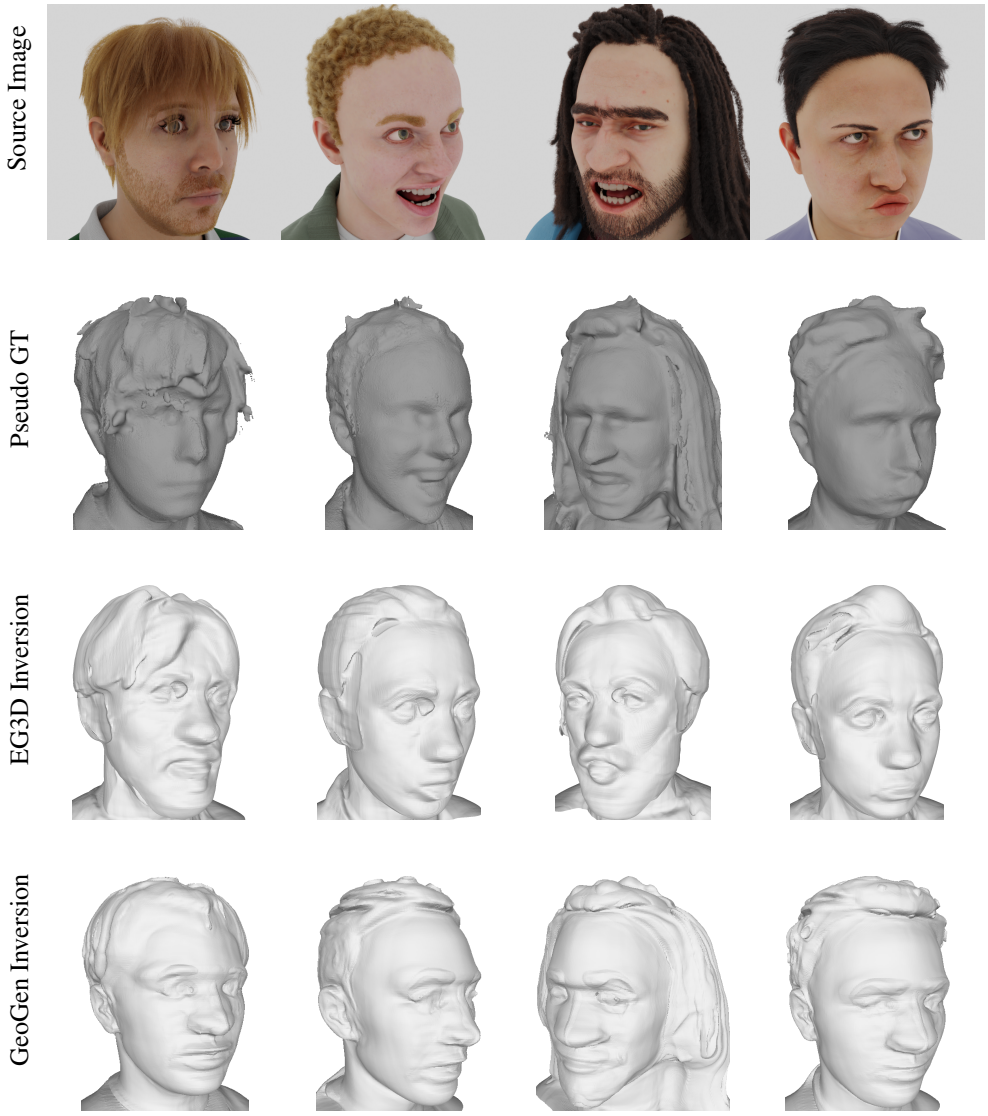

Figure A5. Qualitative inversion results on our synthetic face dataset, focusing on the comparison between the EG3D [3] and GeoGen inversion methods. The corresponding latent source for the source held-out test input image is estimated for GeoGen using GAN inversion, revealing its ability to capture fine details with reduced noise and artifacts. In contrast, the EG3D [3] inversion meshes are observed to have significant artifacts, particularly around the ears, and display noticeable holes in the top regions of the eyes. Our inversion mesh is meticulously compared against pseudo ground truth, and reconstructed using Poisson surface reconstruction from multi-view images, underscoring the superiority of the GeoGen method in terms of fidelity and accuracy. Moreover, our inversion technique exhibits increased precision, contributing to a more authentic representation of the facial structure.

(as shown in the main paper) in reconstructing detailed facial features, including the ears, nose, hair, and eyes.

Additionally, we demonstrate the ability of the GeoGen model in 3D reconstruction on the ShapeNet cars dataset in Figures A4 and Figure A1 where it successfully reproduces granular details on the surface of the cars. We also show inversion results in Figure A3. This distinction is further highlighted by contrasting the rendering qualities of the generated synthetic samples from the EG3D and GeoGen models, displayed in Figure 5, against some ground truth samples. Unlike the EG3D model [3], which exhibits a lack of granular details, our model's implementation of a more advanced SDF network, combined with robust SDF constraints and feature storage within a triplane, yields more precise and refined reconstructions. Thus, our approach consistently and effectively bridges the gap between visual perception and geometric representation, outperforming other techniques in 3D reconstruction fidelity. That is also visible in Fig-

ures A4 and A1 where GeoGen is able to better reconstruct the surface of synthetic faces using a GAN inversion technique [3].

C. Acknowledgments

The authors express their sincere appreciation to Microsoft Research for the provision of GPU clusters containing V100s and P100s. SE's work was supported by the UKRI CDT in Biomedical AI, with additional thanks to the UKRI funds and Microsoft for granting access to cloud services. KK's research received funding from Microsoft Research through the EMEA PhD Scholarship Programme, and he extends his gratitude to NVIDIA Corporation for GPU access provided by NVIDIA's Academic Hardware Grants Program.