GeoGen: Geometry-Aware Generative Modeling via Signed Distance Functions

Supplementary Material

A. Implementation details

The foundation of our model relies on the official implementation of Enhanced Generative 3D Models (EG3D) [2]. 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 [4]. Our model employs the same architecture as StyleGAN2 [7], 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 [2].

A.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 [4]. 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.

A.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

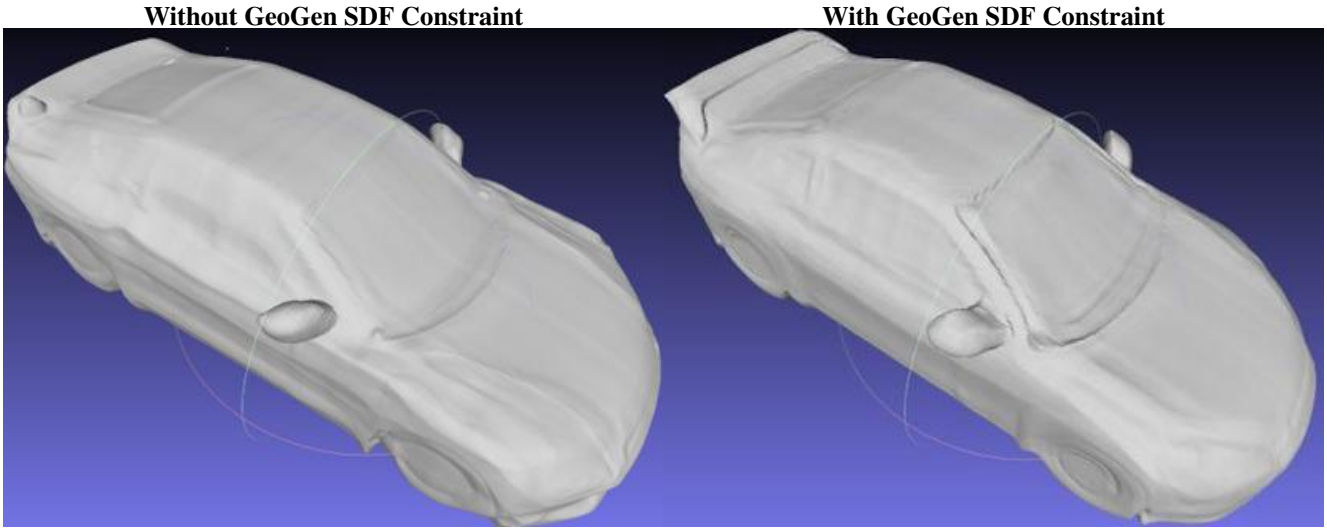

Figure A1. Comparison of models without (left) and with (right) our GeoGen SDF constraint.

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.

A.3. Reconstruction of pseudo ground truth meshes

In our approach to reconstructing pseudo ground truth meshes shown in Figure A5 two methodologies are intricately combined: Planar Prior Assisted PatchMatch Multi-View Stereo (ACMP) [10] and Poisson surface reconstruction [6]. 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 planar compatibility into consideration [10], 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 [6]. 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.

A.4. Results without positional encoder

Understanding Model Collapse in GeoGen without Positional Encoding.This analysis delves into the reasons 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. 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 en-

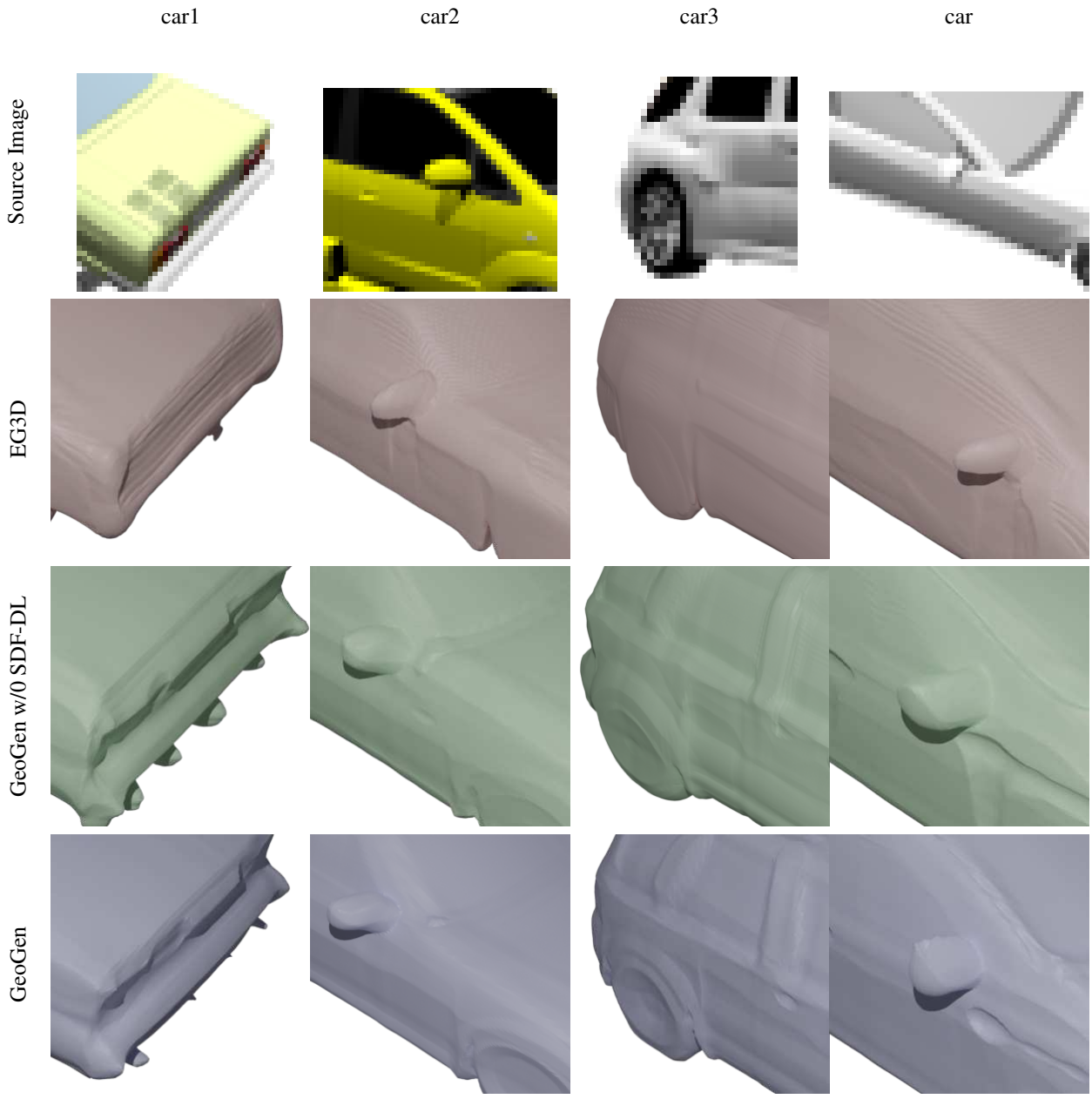

Figure A2. A detailed comparison between EG3D and GeoGen in the context of ShapeNet cars inversion of meshes, emphasizing the differences in the geometric representation and rendering capabilities of both methods. The samples underscore the advanced efficacy of GeoGen in capturing and reconstructing 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 [2] method's rendered meshes reveal a deficiency in portraying granular details, leading to a more approximate and less nuanced depiction of the vehicles.

coding in maintaining the stability and efficacy of models like GeoGen, especially in complex applications involving synthetic human images and 3D rendering.

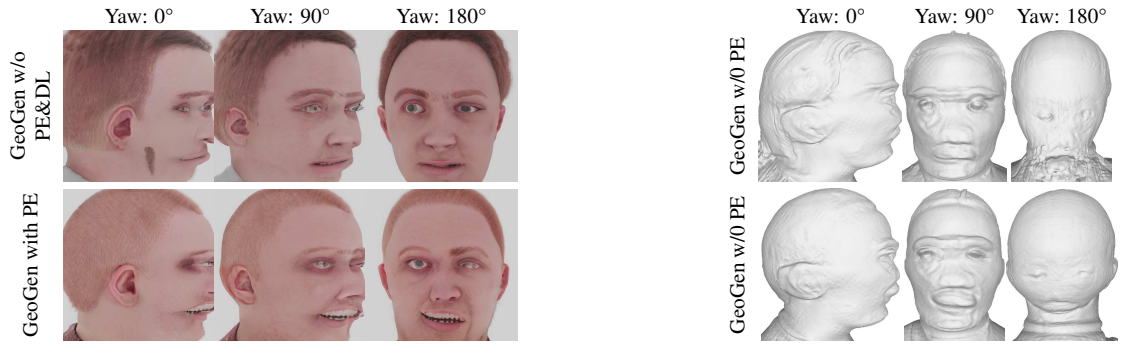

Figure A3. This caption accompanies a series of synthetic images generated by the GeoGen model operating without a positional encoder. The figures on the left illustrate the model's output at different yaw angles, showcasing its ability to render facial features from various perspectives. On the right, 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.

B. Datasets

B.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) [5]. To adapt these images for our purposes, we employed a sophisticated face detection and pose-extraction system [2], 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 [2]. 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 models. 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.

B.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 [3]. 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 [3]. 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 [3].

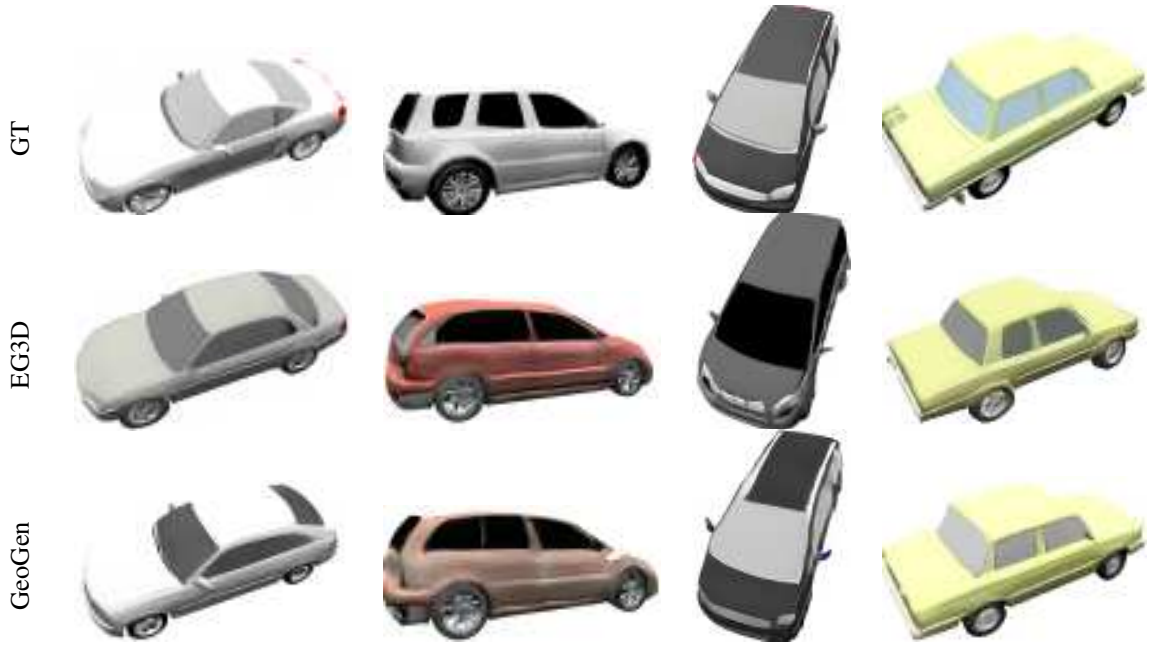

Figure A4. Comparison of EG3D and GeoGen inversion results using held-out images from the ShapeNet Car test set. GeoGen results more closely resemble the input ground truth image (GT).

B.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 [9]. 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.

B.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. [8]. 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.

B.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.

B.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 [5], 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 possible 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) [2] 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.

B.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 al-

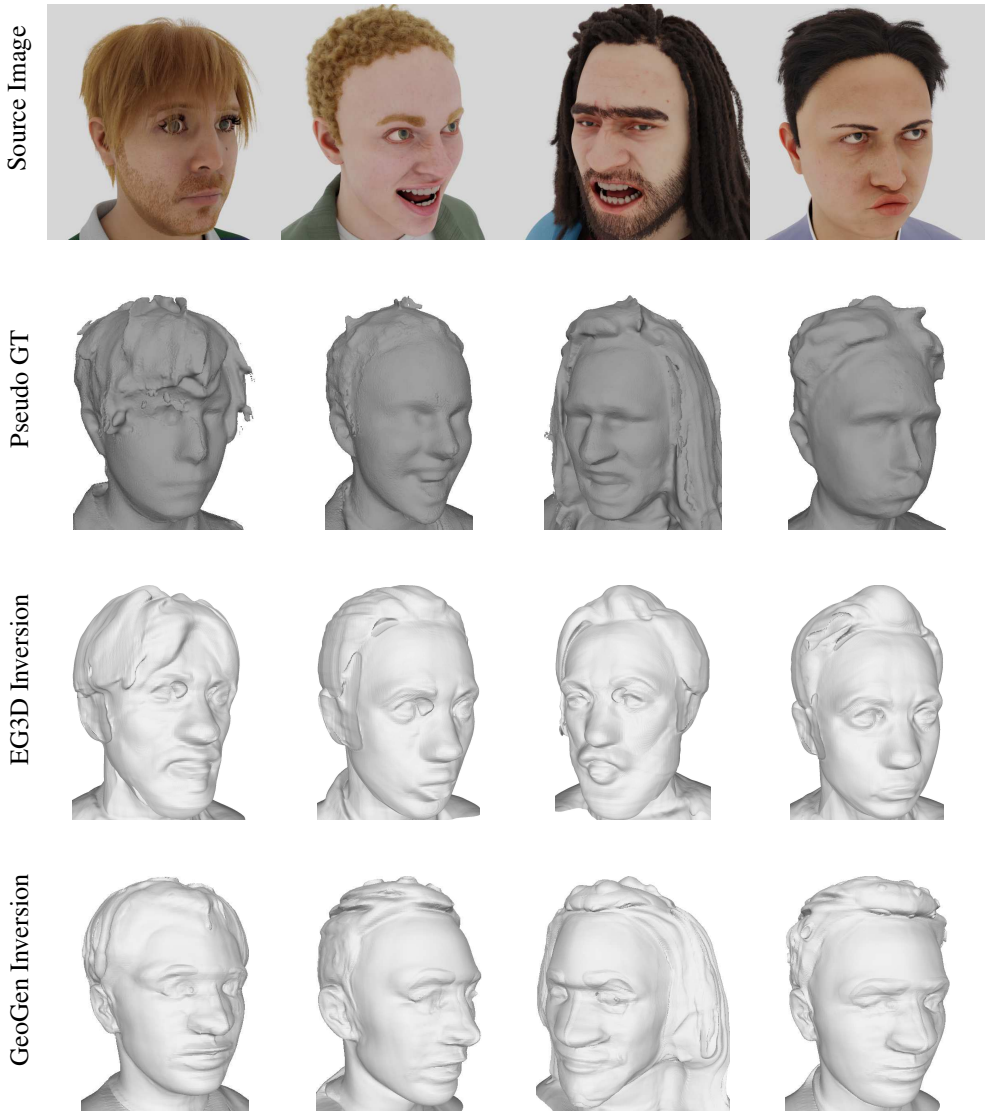

Figure A5. Qualitative inversion results on our synthetic face dataset, focusing on the comparison between the EG3D [2] and GeoGen inversion methods. The corresponding latent source for the source held-out test input image is estimated for GeoGen using GAN inversion, revealing its ability to capture fine details with reduced noise and artifacts. In contrast, the EG3D [2] 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.

lowed 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.

C. Additional qualitative results

In Figure A5 we present a comparison of synthetic human avatar meshes across EG3D [2] 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 (as shown in the main paper) in reconstructing detailed facial features, including the ears, nose, hair, and eyes.

Although StyleSDF [7] also employs an SDF network, it falls short in accurately reconstructing geometry due to the absence of features stored in a triplane structure, and lacking the specific SDF constraints that enable learning fine surface details. This absence results in visible noise and inaccuracies in parts of the faces, reflecting a failure to capture the nuanced geometrical complexities.

Additionally, we demonstrate the ability of the GeoGen model in 3D reconstruction on the ShapeNet cars dataset in Figure A2 and Figure A1 where it successfully reproduces granular details on the surface of the cars. 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 [2], 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 Figure A2 and Figure A1 where GeoGen is able to better reconstruct the surface of synthetic faces using a GAN inversion technique [2].

D. 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.

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. 5

[2] Eric Chan. Efficient geometry aware 3d network. *Computer Vision and Pattern Recognition*, 2022. 1, 3, 4, 6, 7, 8

[3] 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. 4

[4] 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

[5] 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. 4, 6

[6] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Eurographics symposium on Geometry processing*, 2006. 2

[7] 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, 8

[8] E. Tov, J. Doe, and A. Smith. Pivotal tuning inversion. *Computer Graphics Forum*, 2023. 5

[9] 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. 5

[10] Qingshan Xu and Wenbing Tao. Planar prior assisted patchmatch multi-view stereo. In *AAAI Conference on Artificial Intelligence*, 2020. 2