CMSC733: Final project
Semantic-aware Smoothness Regularization for 3D Generative Model

Quynh Phung **Chuong Huynh**
Department of Computer Science
University of Maryland
College Park
{quynhpt,chuonghm}@umd.edu

Abstract

3D-aware GANs make a huge progress in not only generating realistic images but also rendering detailed and high quality 3D shapes. However, the generated shapes still contain a lot of artifacts because of the lack of 3D information in training. We propose a new shape prior named Semantic-aware Smoothness Regularization to improve human portrait shapes generated by those models. The work shows the effectiveness of both image generation and 3D rendering in quantitative and qualitative measurements.

1 Introduction

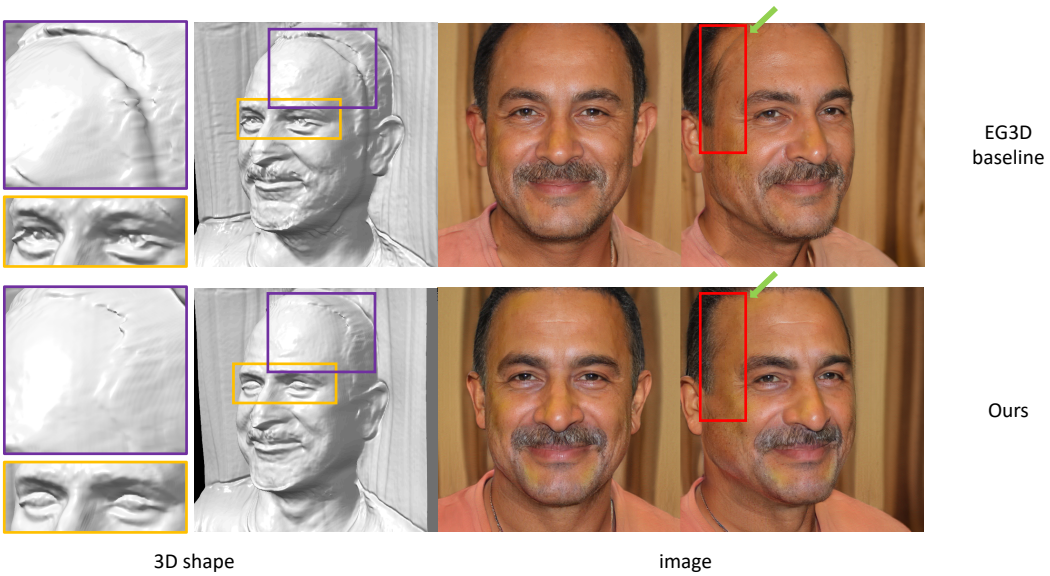

Figure 1: Our proposed method improves the accuracy of 3D shape rendered by the state-of-the-art generative model without losing the detailed texture. While artifacts like trenches in the head or indented eyes exist in the baseline, our work produces more realistic images and their corresponding 3D shape.

CMSC 733: Computer Processing of Pictorial Information (Fall 2022), .

Recent generative adversarial networks (GANs) have achieved huge success with the capability of high-resolution and realistic 2D image generation [1, 2]. However, there is a lack of work on 3D scenes synthesized because of the difficulty in capturing high-definition 3D datasets.

There are a few recent works in 3D-aware GANs that implicitly learn 3D shapes without the requirements of geometrical supervision or multi-view datasets. Those works open a new direction to synthesize new views of objects with high consistency. In rendering techniques, there are mainly explicit voxel grids [3, 4] and neural implicit representation [5, 6, 7], which take part in the solutions for view synthesis. Thanks to the development of neural rendering in recent years [8], adversarial generative models can now generate 3D views easier with the collaboration of rendering modules. However, the high computational cost in memory and speed is still the problem preventing 3D-aware GANs from generating high-definition output.

EG3D [9] is introduced to tackle the problem of expensive 3D shape generation. The proposed hybrid explicit and implicit 3D representation increases computational efficiency and successfully generates high-resolution and realistic images as well as detailed 3D shapes. However, because of no 3D information being used in training the 3D shapes, there are artifacts that exist in the images and shapes generated. Figure 1 illustrates the artifacts in outputs generated by EG3D which are recognizable in both images and 3D shapes. While the trenches along the edge of the faces (a.k.a "seam" artifacts, highlighting in purple in the figure) have been mentioned in recent works [10, 11], the eyes should be in globe-shaped to improve the realistic of 3D representations.

With the idea of improving existing artifacts through prior geometrical knowledge, we propose additional shape constraints to improve the baseline outputs. In this project, we focus on 3D facial generation, and in the future, extensions to any objects will be explored. From our observations, facial 3D shapes need to be smooth in most regions, particularly the skin. That hypothesis is also true in the eyes to make that region nearly globe-shaped. Our proposed constraints help the baseline model achieve better geometrical correctness when reducing artifacts in both types of outputs. The result of the similar generated image is shown in Figure 1.

The contribution of our work includes:

- Propose a smoothness regularization to limit the depth differences between neighbor locations in 3D shape generation.
- Corporate semantic segmentation information to constraints smoothness to regions of interest (skin, eye).

2 Related works

Generative Adversarial Networks (GANs). 2D GANs such as StyleGANv2 [1], BigGAN [12] are capable of generating high-resolution and realistic images. Based on these generators, substantial work [13, 14] manipulates generated images by investigating the latent space direction. Besides, image-to-image translation [15] and image editing [16] have significant progress. However, these 2D GANs lack 3D information, which leads to inconsistency in terms of geometry.

3D representation. There are two main 3D representations: explicit representations, including point cloud, voxel grids, and implicit representations or coordinate networks. While explicit ways [17, 18] are fast to evaluate, it consumes a large memory and is very difficult to scale to high-resolutions. Implicit methods [8, 19] are often implemented with neural networks, so it is slow to evaluate but very memory efficient. There are many works [20, 21] proposing a hybrid approach of implicit and explicit representations which leverage the speed of explicit and the memory efficiency of implicit methods.

3D-aware GANs. 3D-aware GANs have received a great deal of research attention in recent years. Chan et al. [22] proposed an implicit generative model to generate 3D information from 2D images. However, this method is limited to low-quality of generated images and 3D shapes. To improve surface quality, ShadeGAN [23] used a shading-guided generative model, which constrained lighting properties in 3D shape. However, the high resolution or computational cost was still a problem. To reduce the cost of implicit or explicit 3D representation, Eric et al. [9] proposed a hybrid procedure that decomposes 3D volume into three planes. The same idea proposed by Roy et al. [24] samples

implicit 3D representation in rendering without degrading the overall performance. However, until now, the detailed 3D shape without artifacts is still an open problem.

Shape prior. Previously, shape priors are often used in many 3D problems. The paper [25] relies on the optimization of 3D smoothness priors for reconstructing a closed continuous surface from multiple images and silhouettes. Many depth map methods [26, 27, 28] adopt smoothness (regularization) constraints. The paper [29] uses the higher-level shape prior to constraint 3D shapes. This paper utilizes the 2D semantic segmentation in images such as walls, and ground, to apply regulation and surface normal in each region of images. Inspired by previous work, we find that 3D-aware GANs only use 2D images to estimate 3D information without any ground truth, which can lead to the risk of shape-appearance ambiguity. Therefore, it is possible to improve the surface quality by encoding smoothness prior to 3D shapes.

Facial semantic prior. Since the CelebAHQ-Mask [30] was published as a high-quality semantic segmentation for portraits, there are many works using semantic information to generate and manipulate high-quality facial images. While [10, 31] uses semantic mask as the input and supervises with the semantic label in intermediate stages, [32, 33] add pooled semantic features after each convolution block in the generative model.

3 Semantic-aware Smoothness Regularization

We propose a new loss function to improve the smoothness of the 3D shape surface without losing the detailed texture. The proposed method is experimented with facial images and will be extended to other categories in the future. In the following sections, we will explain the added regularization term and customized versions for skin and eyes on generated facial 3D shapes.

3.1 Smoothness regularization (SR)

Let $D \in \mathbb{R}^{h \times w}$ be the estimated depth of generated images by 3D GAN, we assume that the D will be smooth in some regions of interest. In other words, the differences between the depth of each spatial location and their neighbors should be small. Let $k \times k$ (k is odd) be the size of the window where we want to minimize the depth of the center location, the smoothness regularization at $\mathcal{L}_{SR}^{x,y}$ is a sum squared distance of location (x, y) to each neighbor:

$$\mathcal{L}_{SR}^{x,y} = \sum_{i=-k/2}^{k/2} \sum_{j=-k/2}^{k/2} (D_{x+i,y+j} - D_{x,y})^2 \tag{1}$$

where $D_{x,y}$ is the depth value at the location x, y in D.

In practice, the above regularization function can be implemented as $k^2 - 1$ convolution kernels:

$$\mathcal{L}_{SR} = \sum_{i=1}^{k} \sum_{j=1}^{k} \begin{cases} 0, & \text{if } i = j = \lceil k/2 \rceil \\ (f_{i,j} * D)^2, & \text{otherwise} \end{cases} \tag{2}$$

where $\mathcal{L}_{SR} \in \mathbb{R}^{h \times w}$ and $f_{i,j}$ is a convolutional kernel that has one at center $(\lceil k/2 \rceil, \lceil k/2 \rceil)$ and -1 at (i, j). In our experiments, we choose $k = 3$

3.2 Semantic-aware smoothness regularization (SSR)

The assumption of smoothness in Sec. 3.1 is not correct for any spatial locations in D. There are some regions where each location has different depth value such as hair, and background in person portrait images. Semantic segmentation masks are used to constrain the smoothness assumption to some regions of interest (RoI). In this project, we consider two regions in the human face having smoothness properties which are the skin and eyes. Let $S_{RoI} \in [0, 1]^{h \times w}$ is the RoI segmentation mask, we want to compute SSR regularization by:

$$\mathcal{L}_{SSR}^{x,y} = \begin{cases} \mathcal{L}_{SR}^{x,y}, & \text{if } e(S_{RoI})^{x,y} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where e is the erosion function with kernel size k_e to remove boundary regions where the smoothness hypothesis does not satisfy.

3.3 Skin smoothness

We compute $\mathcal{L}_{skin}$ based on $\mathcal{L}_{SSR}$ with S_{RoI} as the skin region in the segmentation mask and $k_e = 7$. However, in our experiments, computing the average value on all skin locations in this large region can make the loss become very small. The reason is that there are many locations has small regularization values and should be ignored in the optimization step. Inspired by prior works in online hard mining [34], we compute the average of the top-p locations with the highest regularization values as the objective function in each optimization iteration. In our experiments, we choose $p = 70\%$.

3.4 Eye smoothness

The eye smoothness $\mathcal{L}_{eye}$ is computed with eye region as S_{RoI}, $k_e = 3$, and no hard-mining applied. However, the smoothness assumption is only true in eye regions when the pose is nearly frontal. Eye regions in faces without frontal poses can have sub-regions with large regularization values, and giving a penalty there can cause unrealistic outputs in both images and 3D shapes.

3.5 Loss function

The final smoothness loss $\mathcal{L}_s$ is the weighted sum of skin and eye smoothness:

$$\mathcal{L}_s = \mathcal{L}_{skin} + \alpha\mathcal{L}_{eye} \tag{4}$$

In our experiment, we choose $\alpha = 0.8$.

4 Experiments

4.1 Dataset

We compare the performance of SSR on the task of unconditional 3D-aware generation with FFHQ [1] dataset. FFHQ is a real-world human face dataset with around 150,000 high-resolution images Flickr. Poses and camera parameters are estimated by [35].

4.2 Implementation details

We do our experiments on the public source of the baseline EG3D [9]. Because of the limitation in time and computational resources, the results are only finetuned on published pre-trained weights of the baseline. All experiments are finetuned in 500 iterations with a batch size of 32 on 8 A100 GPUs in 8 hours. The learning rate used is half of the original ones, particularly 0.001 for the discriminator and 0.00125 for the generator. The finetuning uses the resolution of 128^2 for image generation and neural rendering for the best smoothness optimization.

To supervise the smoothness regularization, we use the public BiSeNet weights [36] on CelebAHQ-Mask to generate the semantic segmentation mask. There are three categories used in the smoothness loss, namely, skin, left, and right eye. The visualization of those regions can be seen in Figure 2.

4.3 Quantitative results

Besides improving the rendered 3D shapes, the added regularization should not degrade the performance of generating images. We use Frechet Inception Distance (FID) [37] and Kernel Inception Distance (KID) [38] on 50,000 generated images to evaluate the performance of models.

The Table 1 shows the performance of the baseline model and different settings of our proposed regularization. The baseline model is the public version of EG3D while finetuned is the baseline added R1 regularization [39] mentioned in their supplementary. The two SSR losses are added to our settings in the last rows. The best FID score yields when both SSR on skin and eye regions are used. Besides, KID achieves lower scores when adding those regularization constraints.

Figure 2: Semantic segmentation mask of EG3D's generated images predicted by the pre-trained model on CelebAHQ-Mask.

Table 1: Adding SSR regularizations not only improve the smoothness of rendered 3D shapes but also the FID and KID scores in generated images. (lower is better)

Model	**R1 Regularization**	**Skin SSR**	**Eye SSR**	**FID**	**KID** $\times 100$
Baseline				4.30	0.139
Finetuned	✓			3.88	0.121
Ours	✓	✓		3.81	**0.102**
Ours	✓	✓	✓	**3.62**	0.119

4.4 Qualitative results

Since there are no suitable quantitative metrics to compare the quality of generated 3D shapes, we conducted a user study on ten different shapes and their corresponding generated images. The result is shown in Table 2 of 20 independent responses from our classmates. More than 75% of answers choosing results of our model.

Fig. 3 shows images generated by EG3D baseline model, the fine-tuned model, our model with smoothness loss applied in skin areas, and our model with smoothness loss applied in both skin and eye areas. We find that our model with smoothness loss can generate comparable image quality with the baseline model. However, our model outperforms in 3D shapes, which are fewer artifacts with a smooth surface. Overall improvement of generated 3D shape can be seen in Fig. 5. It is obvious that 3D shapes generated from our models are more smooth in the skin area, and eyeballs are convex. Therefore, 3D shapes look more realistic by removing artifacts in the eye, forehead areas, etc. 6. Moreover, EG3D baseline model contains "seam" artifacts of side faces in 3D shape as well as generated images. With smoothness loss, our model can remove these artifacts. Fig. 4 demonstrates the difference between our model and the baseline model.

Table 2: Users prefer 3D shapes results generated by our model with smoothness constraints to the model finetuned with R1 regularization.

Model	**User's choice**
Finetuned	21.5%
Ours	78.5%

Figure 3: Our proposed method is still able to generate realistic and high-resolution images

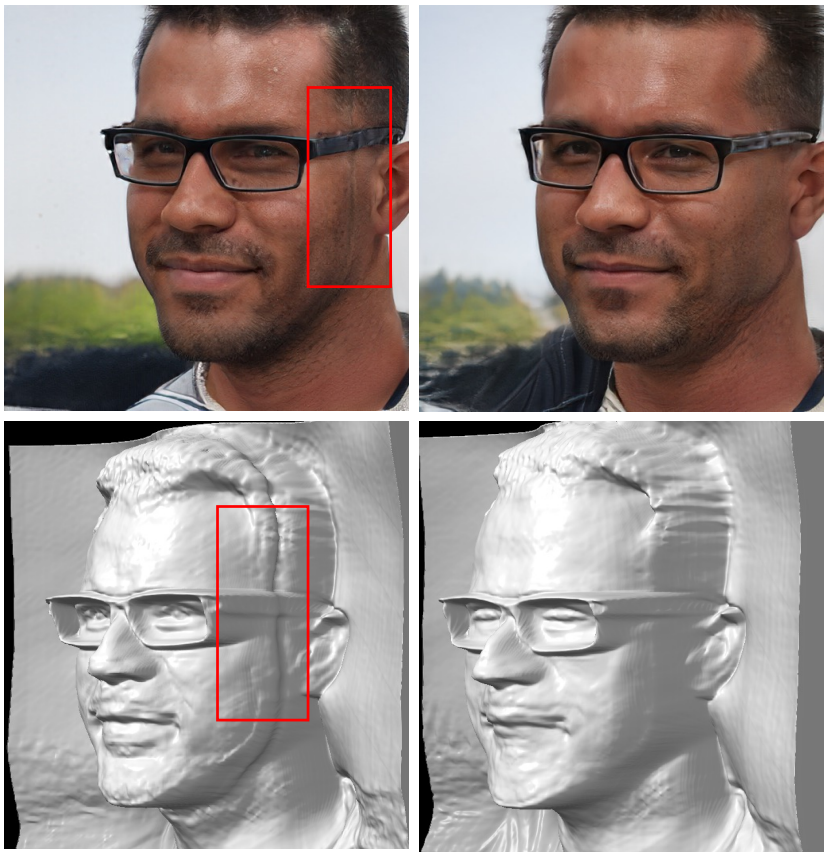

Figure 4: Comparison between images generated from our model and baseline model in terms of side faces.

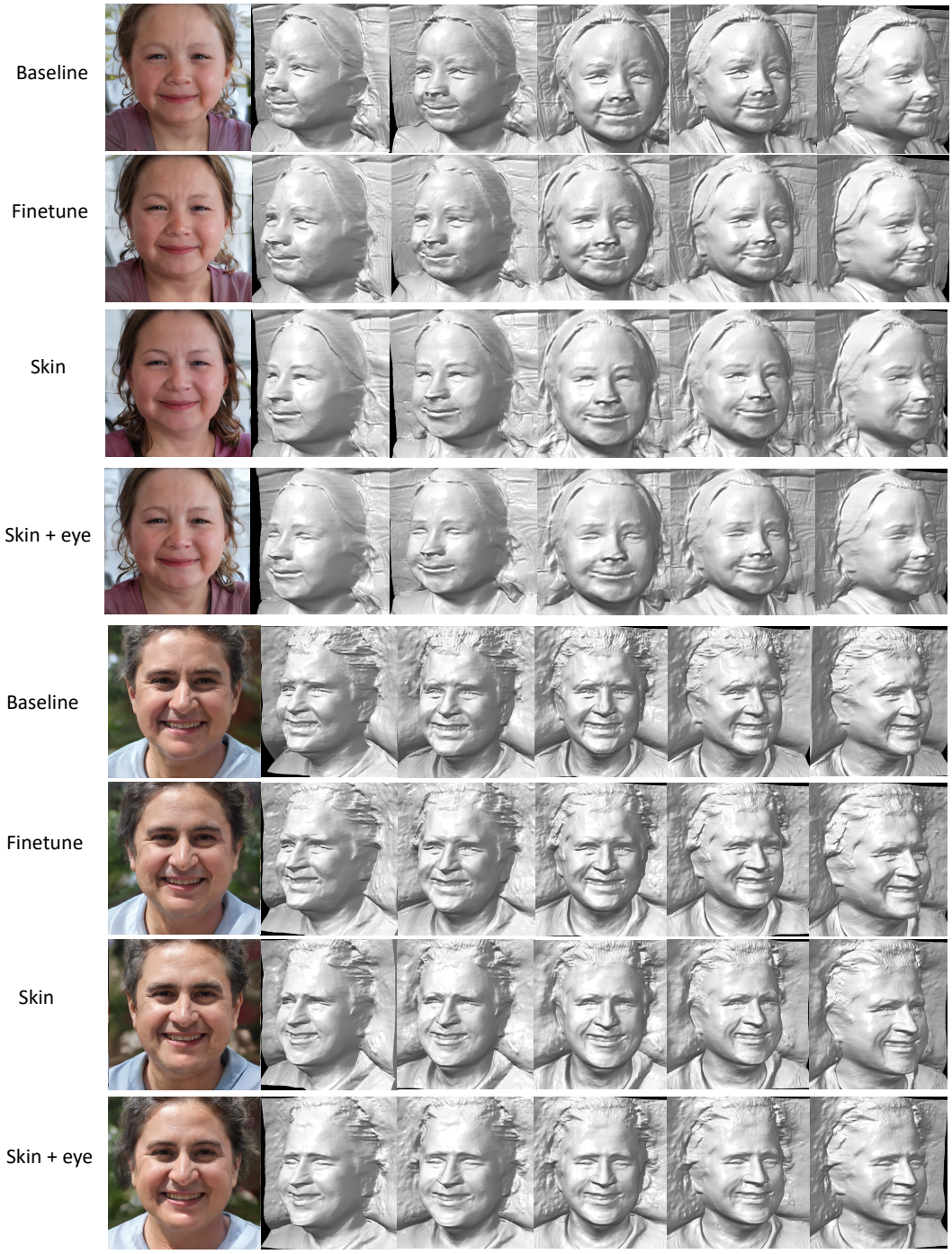

Figure 5: 3D shapes generated from four models: baseline, finetune, smoothness loss in skin and smoothness loss in both skin and eyes.

5 Future works

Due to the limitations in computational resources and time, there are many unfinished works to improve the 3D shape quality. We have planned to constrain the eye regions to be globe-shaped by Gaussian kernel but did not succeed. The symmetry constraints are still unexplored to keep consistency between views. Besides, more experiments on different 3D-aware GANs should be done

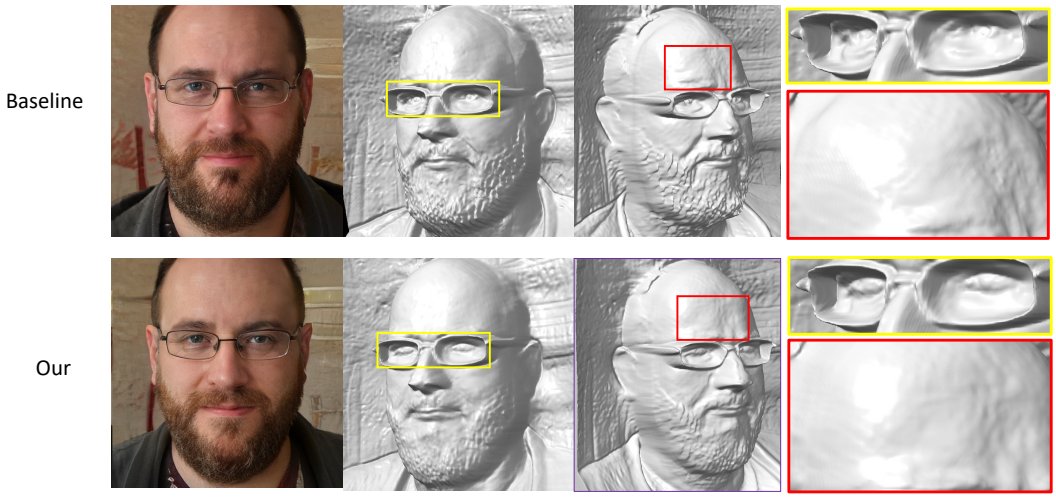

Figure 6: Comparison between images generated from our model and baseline model in terms of eye and forehead area.

to prove the robustness of our proposed method. Finally, we hope that the method can be extended to different objects and datasets.

6 Conclusion

In this project, we propose an efficient semantic-aware geometry constraint for 3D-aware GANs to improve the smoothness and correctness of 3D shapes. Our method helps reduce the artifacts of the EG3D baseline in both images and corresponding shapes. With approximately 0.7 score in FID and 0.02 in KID, our model outperforms the baseline in quantitative metrics. Besides, the proposed regularization also generated better 3D shapes when more than 75% of users choosing results generated by our method.

References

[1] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019.

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

[3] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*, pages 402–411. IEEE, 2017.

[4] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9984–9993, 2019.

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

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13503–13513, June 2022.

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

[8] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

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

[10] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Transactions on Graphics (TOG)*, 41:1–10, 2022.

[11] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021.

[12] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

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

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

[15] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[16] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[17] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2565–2574, 2020.

[18] Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *European Conference on Computer Vision*, pages 608–625. Springer, 2020.

[19] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021.

[20] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W Taylor, and Joshua M Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14304–14313, 2021.

[21] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020.

[22] 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 *CVPR*, pages 5799–5809, 2021.

[23] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. *NeurIPS*, 34:20002–20013, 2021.

[24] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR*, pages 13503–13513, 2022.

[25] Sudipta N Sinha and Marc Pollefeys. Multi-view reconstruction using photo-consistency and exact silhouette constraints: A maximum-flow formulation. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 349–356. IEEE, 2005.

[26] Chaoqiang Zhao, Qiyu Sun, Chongzhen Zhang, Yang Tang, and Feng Qian. Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*, 63(9):1612–1627, 2020.

[27] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[28] Philipp Heise, Sebastian Klose, Brian Jensen, and Alois Knoll. Pm-huber: Patchmatch with huber regularization for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2360–2367, 2013.

[29] Christian Hane, Christopher Zach, Andrea Cohen, Roland Angst, and Marc Pollefeys. Joint 3d scene reconstruction and class segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 97–104, 2013.

[30] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[31] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7850–7859, 2020.

[32] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020.

[33] Yi Wang, Lu Qi, Ying-Cong Chen, Xiangyu Zhang, and Jiaya Jia. Image synthesis via semantic composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13749–13758, 2021.

[34] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[35] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[36] zllrunning. Face parsing with Pytorch. `https://github.com/zllrunning/face-parsing.PyTorch`, 2019.

[37] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *arXiv preprint arXiv:1706.08500*, 12(1), 2017.

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

[39] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.