S3D: Sketch-Driven 3D Model Generation

Hail Song[1*] Wonsik Shin[2*] Naeun Lee[2] Soomin Chung[2] Nojun Kwak[2†] Woontack Woo[1†]
[1]KAIST [2]Seoul National University
{hail96, wwoo}@kaist.ac.kr {wonsikshin, better__62, soomin200, nojunk}@snu.ac.kr

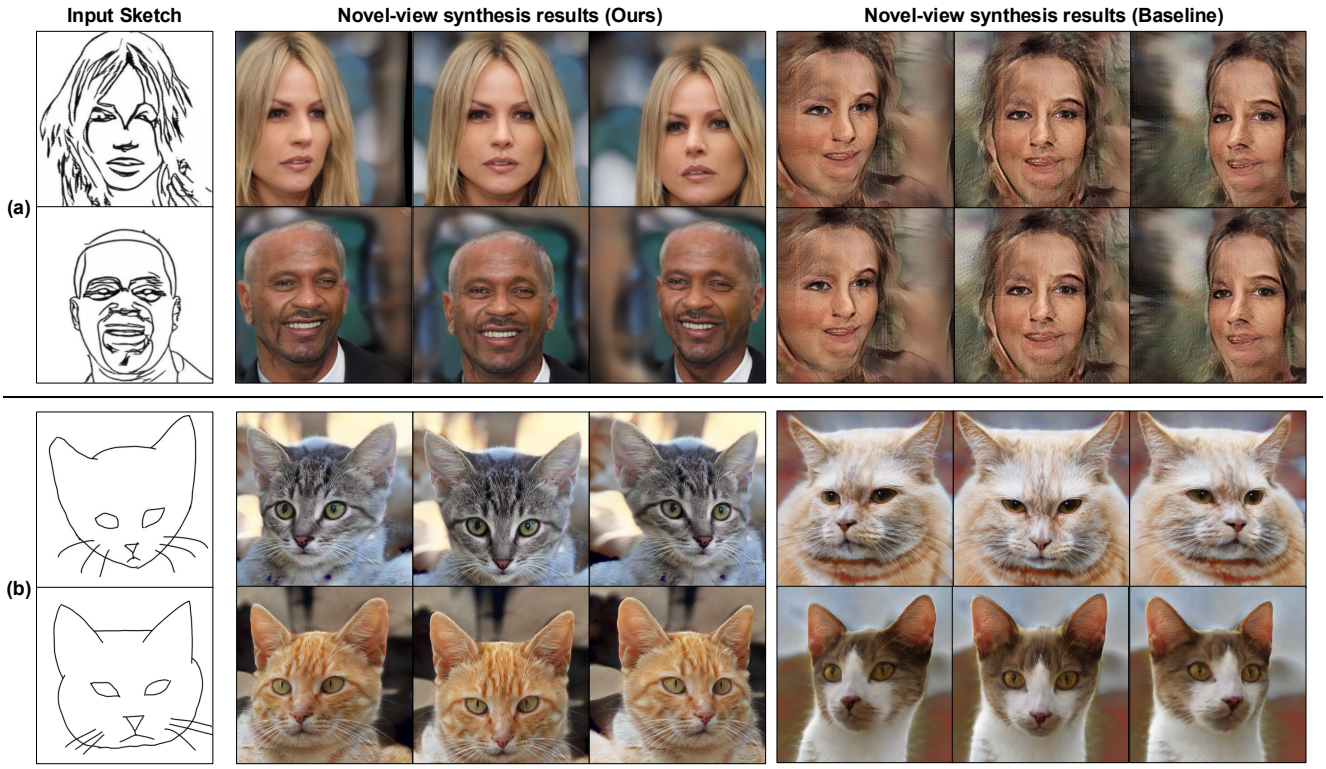

Figure 1. **Comparison between our model and the baseline method.** (a) Given a single sketch image as input, **S3D** (Ours) generates a 3D human face that can be rendered from novel viewpoints. In contrast, the baseline method fails to reconstruct plausible 3D geometry, instead producing nearly identical structures from distinct human sketches. (b) While the baseline often produces models with misaligned facial outlines and unnatural neck contours, S3D generates 3D cat models with high-fidelity textures and more coherent shapes overall.

Abstract

*Generating high-quality 3D models from 2D sketches is a challenging task due to the inherent ambiguity and sparsity of sketch data. In this paper, we present **S3D**, a novel framework that converts simple hand-drawn sketches into detailed 3D models. Our method utilizes a U-Net-based encoder-decoder architecture to convert sketches into face segmentation masks, which are then used to generate a 3D representation that can be rendered from novel views. To ensure robust consistency between the sketch domain and the 3D output, we introduce a novel style-alignment loss that aligns the U-Net bottleneck features with the initial encoder outputs of the 3D generation module, significantly enhancing reconstruction fidelity. To further enhance the network's robustness, we apply augmentation techniques to the sketch dataset. This streamlined framework demonstrates the effectiveness of S3D in generating high-quality 3D models from sketch inputs. The source code for this project is publicly available at https://github.com/hailsong/S3D.*

*Equal contribution.
†Corresponding author.

1. Introduction

The increasing demand for 3D content generation in industries such as film, gaming, and virtual reality has accelerated the development of neural rendering techniques. Neural rendering has been extensively utilized not only for novel view synthesis but also for 3D model generation from visual data. 3D model generation using such techniques often requires multi-view images, which can be impractical in many real-world scenarios. This limitation has led to studies that aim to reconstruct 3D models from a single image input [4]. However, in many practical cases of 3D content creation, concrete image references are simply unavailable, whereas only high level conceptual descriptions exist. As a result, research has shifted toward utilizing more abstract visual inputs—such as 2D segmentation maps—for 3D model generation [7]. Despite these advancements, generating accurate 3D representations from highly abstract inputs—such as hand-drawn sketches—remains a significant challenge.

To address these challenges, we propose **S3D**, a novel end-to-end pipeline for **S**ketch-to-**3D** model generation. Leveraging a divide-and-conquer strategy, our method utilizes a U-Net-based encoder-decoder architecture to convert 2D sketch inputs into segmentation masks, which are then processed by a mask-to-3D module to generate high-fidelity 3D facial models from a single segmentation mask. To enhance the robustness and accuracy of the reconstruction, we employ a combination of Cross-entropy loss and Dice loss. Furthermore, leveraging the prior knowledge embedded in pre-trained models, we introduce a novel loss function that aligns the style vector from the U-Net bottleneck with the initial encoder output of the mask-to-3D module.

Unlike previous approaches, which often struggled with sketch-based 3D face generation due to the complexity of human facial structures and the limited geometric cues in sketches, our model demonstrates the ability to generate high-fidelity 3D representations of both human and cat faces from simple sketch inputs. By enabling accurate 3D face reconstruction from such minimal input, it dramatically reduces the cost and time required for traditional 3D modeling, while unlocking innovative use cases such as efficient forensic facial montages and personalized avatar creation.

In summary, our key contributions are as follows:

- We propose S3D, an end-to-end pipeline for sketch-based 3D face model generation. We integrate a U-Net and a tri-plane-based 3D model generation network to achieve high-quality outputs.
- We introduce a novel style-alignment loss and augmentation strategies to enhance the consistency between sketch-based segmentation masks and 3D reconstructions.
- We enable a new task of generating 3D human faces directly from facial sketches, which was previously unachievable.

2. Related Work

2.1. 3D Neural Representation

Neural rendering techniques, such as Neural Radiance Fields (NeRF) [18], leverage neural networks to represent and reconstruct 3D scenes continuously. NeRF learns a 3D scene representation by training on hundreds of images captured from varying camera transformations, enabling the generation of high-quality 3D renderings and facilitating tasks like novel view synthesis. Recently, advancements in few-shot NeRF [22, 29] techniques have emerged, aiming to achieve comparable performance while reducing the number of input images required for training.

However, despite significant advances in neural 3D reconstruction strategies, these methods still require multiple images captured from different viewpoints under complex setup conditions. This can limit the ease of converting real-world objects or scenes into 3D representations.

2.2. 3D Generation Using Neural Rendering

Unlike methods that construct 3D models or avatars directly from visual information [1, 2, 10, 16, 21, 23, 24, 27, 31], generative approaches operate in a latent space, where they must infer a diverse range of plausible outcomes—posing a fundamentally different challenge. For basic 3D asset generation, recent studies have focused on inferring 3D representations from a single image using various image-to-3D translation methods. Many of these approaches integrate various 3D representations directly into the learning pipeline, employing techniques such as voxel grids [12], voxelized 3D features [19], and 3D morphable models [26]. Other approaches such as StyleNerf [11] leverage NeRF-based models to generate 3D scenes.

Similarly, EG3D [4] introduces a semantic style vector, enabling the generation of high-quality 3D representations using tri-plane-based feature map. Since then, the tri-plane representation has gained widespread adoption and has been widely applied and further developed in numerous 3D generation studies [2, 5, 8, 28, 30]. Among recent approaches, pix2pix3D [7] and SketchFaceNerf [9] utilize sketches or segmentation masks for 3D modeling. However, due to the abstract and ambiguous nature of sketches, these methods exhibit limited generation performance, particularly when dealing with complex structures such as human faces.

3. Methods

We propose **S3D**, a method for **S**ketch-to-**3D** generation by integrating a U-Net-based [20] sketch-to-mask module with tri-plane-based 3D model generation module.

Our method converts sketches into segmentation masks via a sketch-to-mask module, which are then passed to a mask-to-3D module to generate 3D faces from a single segmentation mask. The following sections describe

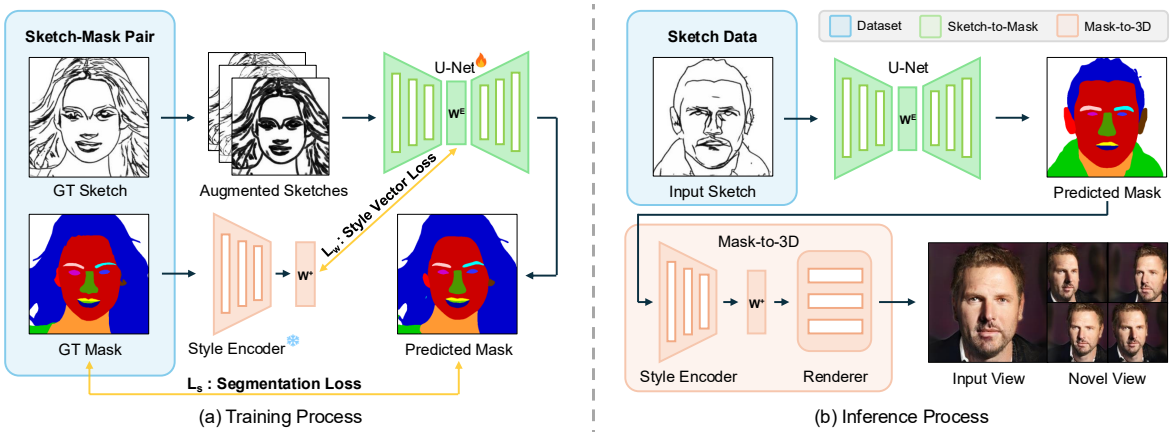

Figure 2. **Pipeline overview of S3D.** Our method combines a U-Net-based sketch-to-mask module with mask-to-3D module to generate 3D models. **(a) Training:** The U-Net is trained with Style Vector loss and Segmentation Loss to align latent features and improve mask accuracy. **(b) Inference:** A sketch is translated into a mask, which is then used by mask-to-3D module to synthesize 3D models.

the sketch-to-mask module, the mask-to-3D module, and the associated training losses. Figure 2 illustrates the full pipeline of S3D.

3.1. Sketch-to-Mask Module

The sketch-to-mask module adopts a CNN-based U-Net architecture that directly maps 512 × 512 input sketches to pixel-wise segmentation masks. Specifically, it comprises seven encoder-decoder pairs, with a bottleneck feature vector of size 7 × 512 which exactly matches the dimensionality of the style vector used in mask-to-3D module. This design choice ensures that the network operates at the same representational capacity as the downstream 3D module. During training, we apply a Style Vector loss that encourages the bottleneck features to match the style vectors extracted from ground-truth masks, to transfer the semantic priors of pretrained mask-to-3D module into sketch-to-mask module. As a result, our network bridges the domain gap between sketch inputs and segmentation outputs, enabling 3D model generation from a single sketch.

3.2. Mask-to-3D Module

We adopt the Seg2Face and Seg2Cat models from pix2pix3D [7] as our mask-to-3D modules for each task. It is a 3D-aware conditional generative model conditioned on 2D label maps, such as segmentation or edge maps. By leveraging a tri-plane feature map, pix2pix3D generates volumetric representations of semantic labels and appearance, which can be rendered from arbitrary viewpoints. Given a 2D label map I_s, the framework first encodes the input using a conditional encoder E, which maps the label map I_s and a random latent code $z \sim \mathcal{N}(0, I)$ to a latent style vector w^+:

$$w^+ = E(I_s, z). \tag{1}$$

The style vector w^+ is then used to modulate a hybrid 3D representation parameterized by tri-planes and a multi-layer perceptron (MLP), enabling the generation of 3D outputs. The MLP computes color $c \in \mathbb{R}^3$, density $\sigma \in \mathbb{R}^+$, feature vectors $\phi \in \mathbb{R}^l$, and semantic labels $s \in \mathbb{R}^c$, where c is the number of classes. To generate a 2D output from a novel viewpoint, pix2pix3D employs volumetric rendering:

$$I_c(r) = \sum_{i=1}^{N} \tau_i c_i, \quad I_s(r) = \sum_{i=1}^{N} \tau_i s_i, \tag{2}$$

where τ_i represents the transmittance probability of a photon along the ray r. Subsequently, the rendered images and their corresponding feature vectors ϕ are provided as input to a CNN-based up-sampling module, which produces high-resolution RGB images $\hat{I}_c^+$ and semantic maps $\hat{I}_s^+$.

3.3. Losses

To train the network, we jointly apply a Style Vector loss and a Segmentation Loss. The Style Vector loss enforces alignment between the U-Net's bottleneck representation and the latent style vector w^+ of pix2pix3D, while the Segmentation Loss ensures accurate transformation of sketches into face segmentation masks.

Let w^E denote the bottleneck embedding, and let w^+ be the style vector of pix2pix3D style encoder. We then define the Style Vector loss as the mean squared error (MSE) between these two vectors:

$$\mathcal{L}_{\text{SV}} = \|w^+ - w^E\|_2^2, \tag{3}$$

By minimizing $\mathcal{L}_{\text{SV}}$, we force the sketch-to-mask module to embed sketches into the same style latent space as segmentation masks, thereby enhancing its ability to reconstruct masks faithfully For Segmentation Loss, we employ the Cross-entropy loss and Dice loss [25].

Dataset	Method	FID ↓	KID ↓	FVV ↓
CelebA	pix2pix3D	232.81	0.3142	0.20
	S3D (Ours)	**21.71**	**0.0065**	**0.18**
AFHQ	pix2pix3D	27.36	0.0054	-
	S3D (Ours)	**23.86**	**0.0047**	-

Table 1. **Quantitative Results.** S3D demonstrated superior performance over the edge-to-3D model of pix2pix3D [7] for 3D generation of human and cat faces. The FVV metric was not measured for the AFHQ dataset, as it is specific to multi-view consistency of human faces.

Dataset	Method	mIoU ↑	mAP ↑
CelebA	w/o $\mathcal{L}_{\text{SV}}$	0.692	0.793
	w/ $\mathcal{L}_{\text{SV}}$	**0.698**	**0.823**
AFHQ	w/o $\mathcal{L}_{\text{SV}}$	0.804	0.884
	w/ $\mathcal{L}_{\text{SV}}$	**0.807**	**0.890**

Table 2. **Ablation Study Results.** Evaluation of mIoU and mAP for our S3D framework with and without the Style Vector loss $\mathcal{L}_{\text{SV}}$ on CelebA and AFHQ datasets.

$$\mathcal{L}_{\text{CE}}(y_{i,c}, \hat{y}_{i,c}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{4}$$

$$\mathcal{L}_{\text{Dice}}(y_{i,c}, \hat{y}_{i,c}) = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_{i=1}^{n}\hat{y}_{i,c}y_{i,c}}{\sum_{i=1}^{n}\hat{y}_{i,c}+\sum_{i=1}^{n}y_{i,c}+\epsilon} \tag{5}$$

Where i denotes an individual pixel, n represents the total number of pixels, c refers to a specific segmentation class, and C is the total number of classes. $y_{i,c} \in \{0, 1\}$ indicates the ground truth label of the i-th pixel for class c, while $\hat{y}_{i,c}$ denotes the predicted probability for the same class. A small constant ϵ is added to the denominator for numerical stability.

The overall training loss $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SV}} + \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Dice}}$ is computed by combining Style Vector loss, Cross-Entropy Loss, and Dice Loss.

4. Experiments

Our experimental results show that our model generates high-quality 3D models and achieves superior performance on quantitative metrics. We used the Multi-Modal-CelebA-HQ dataset [14, 15, 17] for human face generation and the AFHQ [6] dataset for training and testing in cat face generation. Through ablation study, we validate our architectural decisions, with each component contributing to overall performance. We employed the edge-to-3D model of pix2pix3D [7] as a baseline for both quantitative and qualitative evaluation. The details of the augmentation strategy are provided in Appendix A.

4.1. Qualitative Results

Figure 1 presents a comparative analysis between our S3D approach and the baseline, highlighting differences in 3D generation quality. For human face generation, S3D successfully synthesizes both ground truth and novel viewpoints from input sketches, whereas pix2pix3D fails to converge on this task. In the case of cat faces, while pix2pix3D produces multi-view outputs, they exhibit lower fidelity and artifacts compared to the sharper, more coherent results achieved by S3D. Additional qualitative results can be found in Appendix C.

4.2. Quantitative Results

Table 1 presents the quantitative results of S3D and pix2pix3D. For the evaluation metrics, FID [13] and KID [3] are employed to assess the fidelity of generated images. FVV Identity [7] is used to quantify multi-view consistency of the reconstructed human face. The results demonstrate that our S3D consistently outperforms the baseline across both datasets.

4.3. Ablation Study

We compare our full model with an ablated variant without the Style Vector loss $\mathcal{L}_{\text{SV}}$. To quantify the contribution of this loss, we evaluate the quality of the output in terms of the performance of the sketch-to-mask module. Our full model achieves superior mask inference results, indicating that the Style Vector loss $\mathcal{L}_{\text{SV}}$ significantly improves the accuracy of semantic mask generation.

5. Conclusion

We present S3D, a one-shot 3D model generation pipeline. Our key contributions include a Style Vector loss that improves consistency between sketch-to-mask module and mask-to-3D module intermediate representations, complemented by segmentation losses for enhanced mask accuracy. Additionally, we found that our data augmentation strategy contributes significantly to the quality and robustness of the generated results. This approach enables the previously unattainable task of generating 3D human faces directly from facial sketches.

Although S3D demonstrates strong performance in 3D face generation, our study reveals limitations. In particular, a performance gap emerges due to limited diversity in sketches with attributes such as long hair. To address this, we plan to expand the dataset and investigate debiasing techniques as part of our future work.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2

[2] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20950–20959, 2023. 2

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 4

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

[5] Xingyu Chen, Yu Deng, and Baoyuan Wang. Mimic3d: Thriving 3d-aware gans via 3d-to-2d imitation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2338–2348. IEEE Computer Society, 2023. 2

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 4

[7] Kangle Deng, Gengshan Yang, Deva Ramanan, and Jun-Yan Zhu. 3d-aware conditional image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4434–4445, 2023. 2, 3, 4

[8] 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*, 35:31841–31854, 2022. 2

[9] Lin Gao, Feng-Lin Liu, Shu-Yu Chen, Kaiwen Jiang, Chun-Peng Li, Yu-Kun Lai, and Hongbo Fu. Sketchfacenerf: Sketch-based facial generation and editing in neural radiance fields. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2023)*, 42(4):159:1–159:17, 2023. 2

[10] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 2

[11] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis, 2021. 2

[12] Philipp Henzler, Niloy Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering, 2021. 2

[13] 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*, 30, 2017. 4

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. 2018. 4

[15] 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. 4

[16] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot 3d neural head avatar. *Advances in Neural Information Processing Systems*, 36:47239–47250, 2023. 2

[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 4

[18] 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. 2

[19] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images, 2019. 2

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 2

[21] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2

[22] Seunghyeon Seo, Yeonjin Chang, and Nojun Kwak. Flipnerf: Flipped reflection rays for few-shot novel view synthesis, 2023. 2

[23] Hail Song. Toward realistic 3d avatar generation with dynamic 3d gaussian splatting for ar/vr communication. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 869–870. IEEE, 2024. 2

[24] Hail Song, Boram Yoon, Woojin Cho, and Woontack Woo. Rc-smpl: Real-time cumulative smpl-based avatar body generation. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 89–98. IEEE, 2023. 2

[25] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*, page 240–248. Springer International Publishing, 2017. 3

[26] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images, 2020. 2

[27] Phong Tran, Egor Zakharov, Long-Nhat Ho, Anh Tuan Tran, Liwen Hu, and Hao Li. Voodoo 3d: volumetric portrait disentanglement for one-shot 3d head reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10336–10348, 2024. 2

[28] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023. 2

[29] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization, 2023. 2

[30] Han Yi, Zhedong Zheng, Xiangyu Xu, and Tat-seng Chua. Progressive text-to-3d generation for automatic 3d prototyping. *arXiv preprint arXiv:2309.14600*, 2023. 2

[31] Zhongyuan Zhao, Zhenyu Bao, Qing Li, Guoping Qiu, and Kanglin Liu. Psavatar: A point-based morphable shape model for real-time head avatar creation with 3d gaussian splatting. *arXiv preprint arXiv:2401.12900*, 2024. 2

S3D: Sketch-Driven 3D Model Generation

Supplementary Material

A. Data Augmentation

We designed a simple yet effective sketch-specific data augmentation strategy to improve generalization of the sketch-to-mask network. The core idea is to simulate structural perturbations in the sketches using morphological operations.

A.1. Augmentation Details

For each sketch input, we randomly apply one of three augmentation strategies: keeping the original sketch (with a probability of 50%), applying dilation (25%), or applying erosion (25%). Dilation is performed using a kernel size of 3, while erosion uses a kernel size of 7. These augmentations are applied uniformly to both human and cat face sketches, introducing beneficial variability to improve robustness during training.

A.2. Effects of Augmentation Strategy

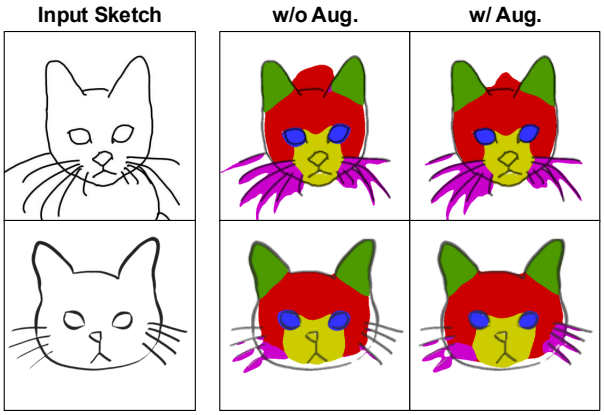

Figure 3. **Effect of data augmentation on training robustness.** Our sketch-specific augmentation strategy enables the generation of diverse sketch styles. Without data augmentation, the model often produces segmentation masks that are smaller than the true facial contours and fails to capture regions such as facial hair. After incorporating augmentation, it generates more robust masks from the input sketches.

Figure 3 illustrates the effect of our sketch-specific augmentation strategy by visualizing predicted segmentation masks before and after augmentation. Because agmented samples exhibit a wide range of stylistic variations, they encourage the network to generalize better and produce more robust mask predictions.

Table 3 illustrates how applying data augmentation affects the performance of the sketch-to-mask network. Although, on the CelebA dataset, the mean Intersection over Union (mIoU) was marginally higher without augmentation, overall augmentation led to performance improvements across both datasets.

Dataset	Method	mIoU ↑	mAP ↑
CelebA	w/o Augmentation	**0.699**	0.818
	w/ Augmentation	0.698	**0.823**
AFHQ	w/o Augmentation	0.804	0.881
	w/ Augmentation	**0.807**	**0.889**

Table 3. **Effect of Data Augmentation.** Comparison of mIoU and mAP for models trained with and without sketch-specific augmentation on CelebA and AFHQ. Data augmentation yields improvements in sketch-to-mask performance.

B. Embedding Space Analysis

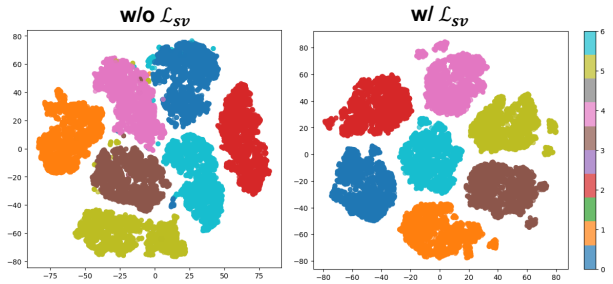

Figure 4. **t-SNE visualization of style embeddings.** Embeddings learned with the style vector loss $\mathcal{L}_{\text{SV}}$ (right) form tight and well-separated clusters, reflecting more structured representation learning. Without $\mathcal{L}_{\text{SV}}$ (left), the distributions are scattered and less semantically aligned.

To further investigate the impact of the style vector loss $\mathcal{L}_{\text{SV}}$, we visualize the learned embedding space using t-SNE in Figure 4. The embeddings produced with $\mathcal{L}_{\text{SV}}$ form tighter and more coherent clusters, indicating that the style vector loss helps guide the network to organize semantic features in a more discriminative manner. Without this constraint, the embeddings are more scattered and inconsistent across samples, particularly within ambiguous sketch inputs.

C. More Qualitative Results

Figure 5 presents additional qualitative results for our proposed S3D model and the baseline. These results underscore the effectiveness of S3D in generating high-fidelity 3D models from sketches on both the CelebA and AFHQ datasets.

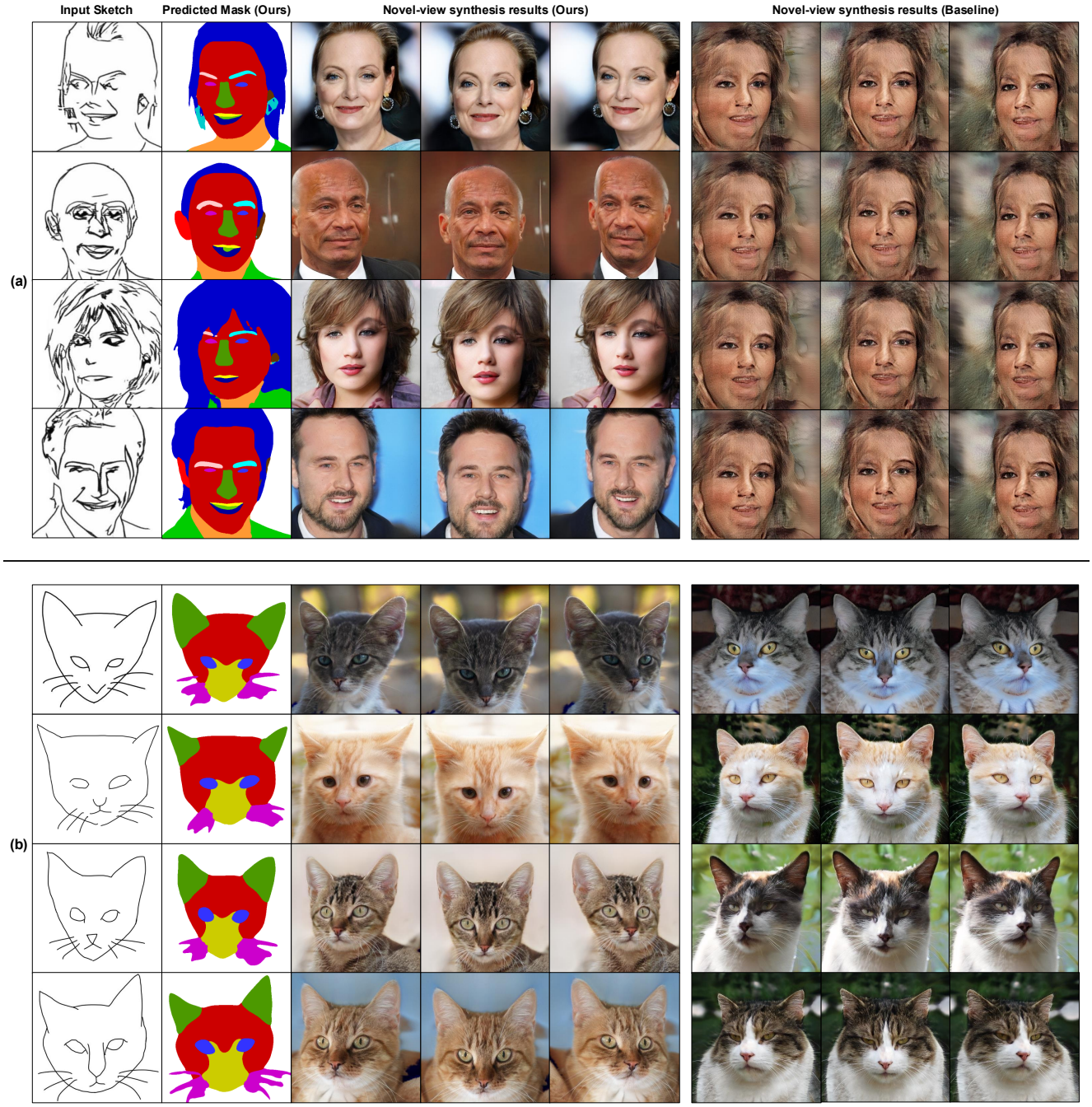

Figure 5. **Additional qualitative comparisons between our S3D method and the baseline.** Across diverse sketch inputs, S3D consistently generates realistic, geometrically coherent 3D models, whereas the baseline produces unnatural structures misaligned with the sketches. (a) On sketch-to-human-face inputs, S3D recovers plausible 3D face geometry, while the baseline fails to form a coherent face. (b) On sketch-to-cat inputs, S3D generates high-quality 3D cat models aligned with the sketches, whereas the baseline yields distorted, misaligned shapes.