

DEEP3DSKETCH: 3D MODELING FROM FREE-HAND SKETCHES WITH VIEW- AND STRUCTURAL-AWARE ADVERSARIAL TRAINING

Tianrun Chen¹, Chenglong Fu², Lanyun Zhu³, Papa Mao⁴, Jia Zhang⁴, Ying Zang^{2*}, Lingyun Sun¹

¹College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang, P.R. China.

²School of Information Engineering, Huzhou University, Huzhou, Zhejiang, P.R. China.

³Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore.

⁴Mafu Laboratory, Moxin (Huzhou) Tech. Co., LTD, Huzhou, Zhejiang, P.R. China.

ABSTRACT

This work aims to investigate the problem of 3D modeling using single free-hand sketches, which is one of the most natural ways we humans express ideas. Although sketch-based 3D modeling can drastically make the 3D modeling process more accessible, the sparsity and ambiguity of sketches bring significant challenges for creating high-fidelity 3D models that reflect the creators' ideas. In this work, we propose a view- and structural-aware deep learning approach, *Deep3DSketch*, which tackles the ambiguity and fully uses sparse information of sketches, emphasizing the structural information. Specifically, we introduced random pose sampling on both 3D shapes and 2D silhouettes, and an adversarial training scheme with an effective progressive discriminator to facilitate learning of the shape structures. Extensive experiments demonstrated the effectiveness of our approach, which outperforms existing methods – with state-of-the-art (SOTA) performance on both synthetic and real datasets.

Index Terms— Sketch, 3D modeling, Computer-Aided Design.

1. INTRODUCTION

The rapid development of portable displays and AR/VR brings tremendous demands for 3D content [1]. Computer-Aided Design (CAD) methods require creators to master sophisticated CAD software commands (*commands knowledge*) and to be able to parse a shape into sequential commands (*strategic knowledge*), which restricts its application in expert users [2, 3]. The restrictions call for the need for alternative methods to open the door to 3D modeling for the masses. In recent years, sketch-based 3D modeling has been recognized as a potential solution, as sketches are one of the most natural ways we humans express ideas. While many works have proposed to perform 3D modeling using sketches, Most existing works either require precise line drawings from multiple views or apply step-by-step workflow with *strategic*

knowledge required [4, 5], which is not friendly for novice users. Other work use template primitives or retrieval-based approaches [6, 7], but lack the customizability.

To mitigate the research gap, we aim to use only one single sketch as the input to generate a complete and high-fidelity 3D model. The approach is designed to fully exploit the human sketches to develop an intuitive and fast 3D modeling approach – generating a high-fidelity 3D model that represents the creators' intention.

However, generating a 3D model from a single sketch is non-trivial. The sparsity and ambiguity of sketches bring significant challenges. Specifically, sketches are sparse because they have only a single view, are mostly abstract, lack fine boundary information when drawing by humans, and, more critically, lack the texture information for depth estimation. This brings large uncertainty when learning 3D shapes. The abstract boundary also makes it hard to interpret, as the same set of strokes may lead to different interpretations in the 3D world, which leads to ambiguity. Existing works [8, 9] have demonstrated that deploying a widely-used auto-encoder as the backbone of the network can only obtain coarse prediction, but is unable to obtain the fine-grained 3D structures.

Facing the challenges, we present our **Deep3DSketch**, a novel and more effective sketch-based modeling approach, which can obtain 3D shapes with fine-grained and reasonable 3D structures. Specifically, we first explicitly learn the view information and use it to condition the generation process to resolve the ambiguity. We then perform random pose sampling to force learning of realistic and high-fidelity 3D shapes independent from the viewpoint. The disentanglement is similar to disentangling "where" and "what" [10]. We also introduce an adversarial training scheme with an effective progressive discriminator that is aware of the geometric structure of the objects via cross-view silhouettes of the 3D model. The discriminator alleviates the uncertainty from the sparsity through more visual clues from different viewpoints, leading to better optimization results. Extensive experiments demonstrated the effectiveness of our approach for generating 3D models with higher fidelity, achieving state-of-the-art (SOTA)

* 02750@zjhu.edu.cn

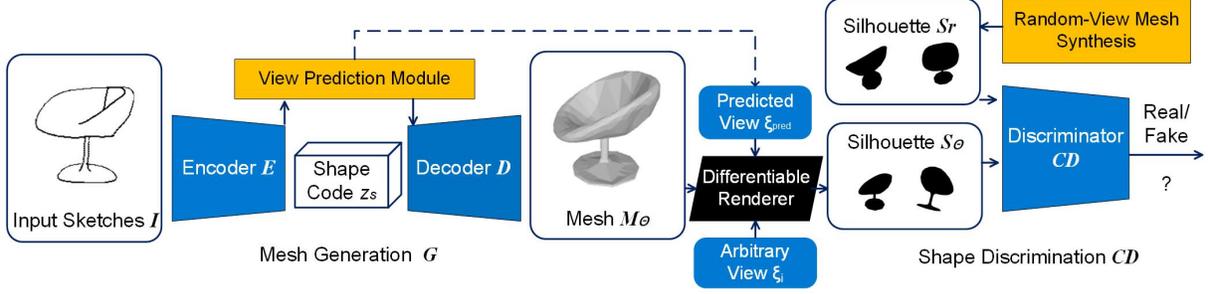


Fig. 1. The structure of Deep3DSketch. View- and structural- aware sketch-based 3D modeling with adversarial training.

performance on both synthetic and real datasets.

2. METHOD

2.1. Preliminary

Given the input binary sketch $I \in \{0, 1\}^{W \times H}$, the goal of the network G is to obtain a mesh $M_\Theta = (V_\Theta, F_\Theta)$, in which V_Θ and F_Θ represents the mesh vertices and facets, and the rendered silhouette $S_\Theta : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ of M_Θ matches with the information from the input sketch I . We use commonly-used encoder-decoder structure as the backbone, an Encoder E is used to obtain a compressed shape code z_s and a Decoder D manipulate z_s to calculate the vertex offsets of a template mesh and deforms it to get the output mesh $M_\Theta = D(z_s)$.

2.2. View-Aware 3D Model Generation

We first introduce extra clues – view information, which can tackle the challenge of ambiguity [9]. As we humans use viewpoint clues to recognize and interpret 3D objects, viewpoint clues are important in single sketch-based 3D modeling, especially in resolving ambiguity. Therefore, we explicitly learn the viewpoint and use the viewpoint information to condition the generation process. We first let the encoder E produce another latent code z_l and input it to the viewpoint prediction module. We implement two fully-connected layers to produce the viewpoint estimation ξ_{pred} , represented by an Euler angle. The viewpoint prediction module is optimized in a fully-supervised manner, with the input of the ground truth viewpoint and supervised by a viewpoint prediction loss \mathcal{L}_v , which adopted MSE loss for predicted and ground truth camera pose, defined as:

$$\mathcal{L}_v = \|V - \hat{V}\|_2 = \|V - D_v(z_v)\|_2 \quad (1)$$

The output viewpoint prediction ξ_{pred} is fed into a differentiable renderer to render silhouette at the given viewpoint for supervision. Specially, we use the mIoU Loss \mathcal{L}_{iou} to measure the similarity between the rendered silhouette S_1 and the

silhouette of the input sketch S_2 :

$$\mathcal{L}_{iou}(S_1, S_2) = 1 - \frac{\|S_1 \otimes S_2\|_1}{\|S_1 \oplus S_2 - S_1 \otimes S_2\|_1} \quad (2)$$

For computational efficiency, we progressively increase the resolutions of silhouettes, forming the multi-scale mIoU loss \mathcal{L}_{sp} , which is represented as

$$\mathcal{L}_{sp} = \sum_{i=1}^N \lambda_{s_i} \mathcal{L}_{iou}^i \quad (3)$$

The predicted viewpoint is also used to guide the generation process. We fed the viewpoint into another two fully-connected layers to produce a view-aware vector representation z_v , and input both z_v and z_s to the Decoder D to produce the M_Θ .

To further condition the generation process with viewpoint constraints, we add a Random-View Mesh Synthesis branch, in which a random viewpoint ξ_{random} is obtained and a mesh $M_{\Theta r}$ is generated following the same manner as mesh generation with ξ_{pred} . The generated $M_{\Theta r}$ with random (fake) viewpoint constraint is regarded as the fake sample, while the generated mesh M_Θ is regarded as the real sample. They together feed into a Shape Discriminator CD , to force the neural network generate meshes under the view-constraint.

2.3. Structural-Aware 3D Model Generation

So far, the supervision of the mesh generation fidelity is from a single rendered silhouette of generated mesh with a given viewpoint as the input. With only 2D input as the supervision, our goal is, however, to obtain complete 3D shapes with fine-grained structural information. A single sketch and the corresponding silhouette can only represent the information at that given viewpoint, but lacks the information from other viewpoints, thus making it hard to obtain detailed structural information. The sparsity of the sketch contributed to the difficulty of obtaining fine-grained structures. Therefore, we propose to have multiple random-view silhouettes. The random pose

sampling aims to force the network learns to generate reasonable 3D fine-structured shapes independent from the view-points. In addition, as many previous works investigated in the realm of shape-from-silhouette, the proposed multi-view silhouettes contain valuable geometric information about the 3D object [11, 12, 13]. In practice, we randomly sample N camera poses $\xi_{1\dots N}$ from camera pose distribution p_ξ . We use a differentiable renderer to render the silhouettes $S_{1\dots N}$ from the mesh M and render the silhouettes $S_r \{1\dots N\}$ from the mesh M_r . The differentiable renderer R is shown in [14]. By introducing the $S_r \{1\dots N\}$, the network is aware of the geometric structure of the objects in cross-view silhouettes when generating the 3D objects, and the discriminator helps to resolve the challenge from the sparsity of sketches by offering more visual clues. The disentanglement is very similar to disentangling "where" and "what" principles in generative models [10], which is proven to be effective in our tasks.

In addition, to fully capture the structural information of the rendered silhouettes, we apply a convolutional progressive growing discriminator CD . Following [15], our discriminator is trained with increasing image resolution and incrementally added new layers to handle the higher resolutions and discriminate fine details. We discovered that such convolutional discriminator design is more effective in capturing local and global structural information to facilitate the generation of high-fidelity 3D shapes, compared to MLP-enabled discriminator for 3D objects. In training, non-saturating GAN loss with R1 regularization is used [16] for better convergence:

$$\mathcal{L}_{sd} = \mathbf{E}_{\mathbf{z}_v \sim p_{z_v}, \xi \sim p_\xi} [f(CD_{\theta_D}(R(M, \xi)))] + \mathbf{E}_{\mathbf{z}_{vr} \sim p_{z_{vr}}, \xi \sim p_\xi} [f(-CD_{\theta_D}(R(M_r, \xi)))] \quad (4)$$

$$\text{where } f(u) = -\log(1 + \exp(-u)) \quad (5)$$

2.4. Domain Adaptation

Due to the lack of large amount of ground truth 3D models and the corresponding 2D sketches, we use synthetic data for training and testing at real-world data, in which the domain gap exists. To make our network generalizable to real hand-draw datasets, we applied domain adaptation (DA) technique

and introduce the DA loss \mathcal{L}_{dd} , as the same in [9].

2.5. Training Details

Loss Function. To make meshes more realistic with higher visual quality, we also use flatten loss and Laplacian smooth loss in [9, 17, 14], represented by \mathcal{L}_r . The overall loss function \mathcal{L} is calculated as the weighted sum of five components:

$$\mathcal{L} = \mathcal{L}_{sp} + \mathcal{L}_r + \lambda_v \mathcal{L}_v + \lambda_{sd} \mathcal{L}_{sd} + \lambda_{dd} \mathcal{L}_{dd} \quad (6)$$

Implementation Details. We use ResNet-18 [18] as the encoder for image feature extraction. The extracted 512-dim feature goes through 2 linear layers with L2-normalization and generates a 512-dim shape code z_s and a 512-dim view code z_v . The rendering module is SoftRas [14], the number of views $N = 3$. Each 3D object is placed with 0 in evaluation and 0 in azimuth angle in the canonical view, with a fixed distance to the camera. We use Adam optimizer with an initial learning rate of 1e-4, and multiply by 0.3 for every 800 epochs. Betas equal 0.9 and 0.999. The total training epochs equal 2000. The model is trained individually with each class. λ_r, λ_{sd} , and λ_{dd} in Equation. 6 equal to 0.1, λ_v and λ_{vr} , equal to 10. When evaluating with the ShapeNet-Sketch dataset, we use domain adaptation on 7 of the classes, which have sufficient amount of sketches in the Sketchy dataset [19] and Tu-Berlin dataset [20]. The domain adaptation is performed by concatenating the average pooling and max pooling results of the image feature map as input, as in [21].

3. EXPERIMENTS

3.1. Datasets

Training the model requires large-scale sketch data with the corresponding 3D models, which is rare in the public domain. Following [9], we use the synthetic data *ShapeNet-synthetic* for training and testing, and the real-world data *ShapeNet-Sketch* to evaluate the method in the wild. ShapeNet-synthetic is the edge map extracted by a canny edge detector provided by Kar et al. [22]. It contains 13 categories of 3D objects from ShapeNet. The ShapeNet-Sketch is a dataset

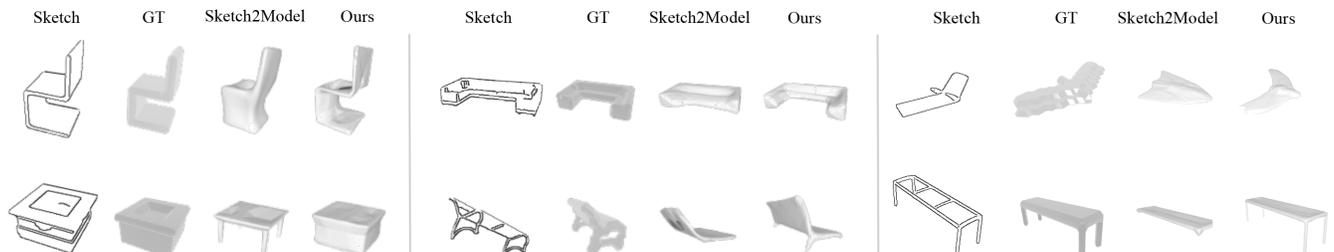


Fig. 2. Qualitative evaluation with existing state-of-the-art. The visualization demonstrated our method’s capability of synthesizing higher fidelity 3D structures.

Table 1. The quantitative evaluation of ShapeNet-Synthetic dataset.

Shapenet-synthetic (Voxel IoU \uparrow)														
	car	sofa	airplane	bench	display	chair	table	telephone	cabinet	loudspeaker	watercraft	lamp	rifle	mean
Retrieval	0.667	0.483	0.513	0.38	0.385	0.346	0.311	0.622	0.518	0.468	0.422	0.325	0.475	0.455
Auto-encoder	0.769	0.613	0.576	0.467	0.541	0.496	0.512	0.706	0.663	0.629	0.556	0.431	0.605	0.582
Sketch2Model (GT Pos)	0.751	0.622	0.624	0.481	0.604	0.522	0.478	0.719	0.701	0.641	0.586	0.472	0.612	0.601
Sketch2Model (Pred Pos)	0.746	0.620	0.618	0.477	0.550	0.515	0.470	0.673	0.667	0.624	0.569	0.463	0.606	0.584
Ours (GT Pos)	0.796	0.651	0.644	0.500	0.612	0.544	0.518	0.738	0.705	0.651	0.595	0.469	0.619	0.618
Ours (Pred Pos)	0.793	0.649	0.641	0.500	0.583	0.541	0.504	0.680	0.683	0.623	0.580	0.465	0.619	0.604

Table 2. The quantitative evaluation of ShapeNet-Sketch dataset.

Shapenet-sketch (Voxel IoU \uparrow)														
	car	sofa	airplane	bench	display	chair	table	telephone	cabinet	loudspeaker	watercraft	lamp	rifle	mean
Retrieval	0.626	0.431	0.411	0.219	0.338	0.238	0.232	0.536	0.431	0.365	0.369	0.223	0.413	0.370
Auto-encoder	0.648	0.534	0.469	0.347	0.472	0.361	0.359	0.537	0.534	0.533	0.456	0.328	0.541	0.372
Sketch2Model (GT Pos)	0.659	0.534	0.487	0.366	0.479	0.393	0.357	0.554	0.568	0.526	0.450	0.338	0.534	0.483
Sketch2Model (Pred Pos)	0.649	0.528	0.479	0.357	0.435	0.383	0.361	0.551	0.547	0.544	0.466	0.336	0.510	0.470
Sketch2Model + DA (GT Pos)	0.679	0.548	0.526	0.367	-	0.398	0.357	-	-	-	-	-	0.535	0.489
Sketch2Model + DA (Pred Pos)	0.659	0.533	0.515	0.362	-	0.385	0.360	-	-	-	-	-	0.511	0.475
Ours (GT Pos)	0.695	0.528	0.502	0.364	0.493	0.389	0.370	0.574	0.563	0.538	0.477	0.334	0.535	0.489
Ours (Pred Pos)	0.683	0.523	0.502	0.364	0.493	0.389	0.370	0.527	0.549	0.509	0.468	0.331	0.535	0.476
Ours + DA (GT Pos)	0.699	0.538	0.517	0.362	-	0.390	0.360	-	-	-	-	-	0.545	0.491
Ours + DA (Pred Pos)	0.692	0.532	0.515	0.360	-	0.382	0.346	-	-	-	-	-	0.545	0.477

collected from real-human drawings. Volunteers with varied drawing skills are asked to draw objects based on the rendered images of 3D objects, with a total number of 1300 sketches and their corresponding 3D shapes.

3.2. Results

The ShapeNet-Synthetic Dataset. We first evaluate the performance on the dataset with the ground truth 3D model. Meshes with the predicted viewpoint (Pred Pos) and the ground truth viewpoint (GT Pos) are trained and evaluated, respectively. We apply common-used 3D reconstruction metrics – voxel IoU to measure the fidelity. The result is shown in Table 3.2. Our method achieves state-of-the-art (SOTA) performance in every category. The quantitative evaluation of our method compared with existing state-of-the-art in Figure 2 further demonstrated the effectiveness of our approach to reconstructing models with higher fidelity in structure.

The ShapeNet-Sketch Dataset. We further evaluate the performance of real-world human drawings. We train the model on ShapeNet-Synthetic dataset and use ShapeNet-Sketch dataset for evaluation. As shown in Table 3.2, our model outperforms the existing state-of-the-art methods in most categories. Our method outperforms the existing method in some categories even without Domain Adaptation (DA).

3.3. Ablation Study

To show the effectiveness of our proposed method, we conducted the ablation study that removes Random Pose Sampling (RPS) for view-awareness. We also remove the progressive Convolutional Discriminator (CD) and use an MLP-based discriminator as in [9]. Our quantitative result (Table 3) and qualitative example (Figure 3) shows removing the RPS and CD will be detrimental to the performance.

Table 3. Quantitative evaluation of ablation study.

Ablation Study. (Numbers inside and outside the parenthesis are IoU on Pred View and GT View, respectively)														
RPS	CD	car	sofa	airplane	bench	display	chair	table						
		0.747 (0.753)	0.624 (0.643)	0.557 (0.565)	0.345 (0.460)	0.457 (0.577)	0.499 (0.508)	0.406 (0.427)						
✓		0.782 (0.773)	0.641 (0.639)	0.644 (0.639)	0.461 (0.485)	0.597 (0.540)	0.543 (0.538)	0.512 (0.477)						
✓	✓	0.796 (0.793)	0.651 (0.649)	0.644 (0.641)	0.500 (0.500)	0.612 (0.583)	0.544 (0.541)	0.518 (0.504)						
RPS	CD	telephone	cabinet	loudspeaker	watercraft	lamp	rifle	mean						
		0.522 (0.705)	0.597 (0.579)	0.584 (0.614)	0.574 (0.575)	0.290 (0.421)	0.500 (0.576)	0.516 (0.569)						
✓		0.734 (0.673)	0.696 (0.645)	0.636 (0.599)	0.585 (0.553)	0.478 (0.471)	0.619 (0.627)	0.608 (0.588)						
✓	✓	0.738 (0.680)	0.705 (0.683)	0.651 (0.623)	0.595 (0.580)	0.469 (0.465)	0.619 (0.619)	0.618 (0.604)						

**Fig. 3. Visualization of Ablation.** The network generates unwanted structures w/o RPS and unrealistic structure w/o CD.

4. CONCLUSION

We propose *Deep3DSketch*, a novel 3D modeling approach that generates 3D models with only a single sketch. For high-fidelity 3D modeling, we disentangle the learning of view and structural learning. We first condition the generation on an explicitly learned viewpoint, then use random pose sampling for viewpoint-independent learning of shapes and fully exploit the geometric information in cross-view silhouettes. We also introduce a progressive convolutional discriminator to better capture the structural information of a 3D mesh at local and global levels. With the alleviated ambiguity and sparsity, we have shown state-of-the-art (SOTA) performance on both real and synthetic data. We believe our method has great potential to revolutionize future 3D modeling pipelines.

5. ACKNOWLEDGEMENT

This work is supported by National Key R&D Program of China (2018AAA0100703). We thank Zejian Li for advice.

6. REFERENCES

- [1] Miao Wang, Xu-Quan Lyu, Yi-Jun Li, and Fang-Lue Zhang, "Vr content creation and exploration with deep learning: A survey," *Computational Visual Media*, vol. 6, no. 1, pp. 3–28, 2020.
- [2] Suresh K Bhavnani, Bonnie E John, and Ulrich Fleming, "The strategic use of cad: An empirically inspired, theory-based course," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 1999, pp. 183–190.
- [3] Ivan Chester, "Teaching for cad expertise," *International Journal of Technology and Design Education*, vol. 17, no. 1, pp. 23–35, 2007.
- [4] Jonathan M Cohen, Lee Markosian, Robert C Zeleznik, John F Hughes, and Ronen Barzel, "An interface for sketching 3d curves," in *Proceedings of the 1999 symposium on Interactive 3D graphics*, 1999, pp. 17–21.
- [5] Congyue Deng, Jiahui Huang, and Yong-Liang Yang, "Interactive modeling of lofted shapes from a single image," *Computational Visual Media*, vol. 6, no. 3, pp. 279–289, 2020.
- [6] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung, "On visual similarity based 3d model retrieval," in *Computer graphics forum*. Wiley Online Library, 2003, vol. 22, pp. 223–232.
- [7] Fang Wang, Le Kang, and Yi Li, "Sketch-based 3d shape retrieval using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1875–1883.
- [8] Benoit Guillard, Edoardo Remelli, Pierre Yvernav, and Pascal Fua, "Sketch2mesh: Reconstructing and editing 3d shapes from sketches," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 13023–13032.
- [9] Song-Hai Zhang, Yuan-Chen Guo, and Qing-Wen Gu, "Sketch2model: View-aware 3d modeling from single free-hand sketches," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6012–6021.
- [10] Xinqi Zhu, Chang Xu, and Dacheng Tao, "Where and what? examining interpretable disentangled representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5861–5870.
- [11] Matheus Gadelha, Rui Wang, and Subhransu Maji, "Shape reconstruction using differentiable projections and deep priors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 22–30.
- [12] Xuyang Hu, Fan Zhu, Li Liu, Jin Xie, Jun Tang, Nian Wang, Fumin Shen, and Ling Shao, "Structure-aware 3d shape synthesis from single-view images.," in *BMVC*, 2018, pp. 230–243.
- [13] Enliang Zheng, Qiang Chen, Xiaochao Yang, and Yuncai Liu, "Robust 3d modeling from silhouette cues," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 1265–1268.
- [14] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li, "Soft rasterizer: A differentiable renderer for image-based 3d reasoning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7708–7717.
- [15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.
- [16] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin, "Which training methods for gans do actually converge?," in *International conference on machine learning*. PMLR, 2018, pp. 3481–3490.
- [17] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada, "Neural 3d mesh renderer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3907–3916.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [19] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays, "The sketchy database: learning to retrieve badly drawn bunnies," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–12, 2016.
- [20] Mathias Eitz, James Hays, and Marc Alexa, "How do humans sketch objects?," *ACM Transactions on graphics (TOG)*, vol. 31, no. 4, pp. 1–10, 2012.
- [21] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [22] Abhishek Kar, Christian Häne, and Jitendra Malik, "Learning a multi-view stereo machine," *Advances in neural information processing systems*, vol. 30, 2017.