RECONSTRUCTING PART-LEVEL 3D MODELS FROM A SINGLE IMAGE

Dingfeng Shi[1], *Yifan Zhao*[1], *Jia Li*[1,2,*]

[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University
[2] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University
{shidingfeng, zhaoyf, jiali}@buaa.edu.cn

ABSTRACT

Understanding an image with 3D representations has been an increasingly attractive topic in computer vision. The state-of-the-art 3D reconstruction methods usually focus on the reconstruction of the holistic object, while missing important part information, which is crucial in robotic interaction and virtual reality applications. To solve this issue, we make the first attempt to reconstruct the 3D models with part-level representations in a unified framework. With the input of the single-view images, we first develop a feature enhancement encoder to incorporate discriminative local features into the feature representation. The local features are selected adaptively by a learnable local awareness module. Then the enhanced local features are fused with the global branch to form the 3D representations. We then develop a 3D part generator to decode the image priors to 3D parts with a 3D focal loss, which enables the representations of small parts. Experimental results indicate that our model generates reliable part-level structures while achieving state-of-the-art performance in object-level recovering.

Index Terms— 3D reconstruction, single-view, part-level

1. INTRODUCTION

Recovering an image into 3D representation is a crucial step to understand and interact with the world. Recent 3D reconstruction approaches [1, 2] have shown promising results in modeling single image into 3D object, which can serve as prerequisites for many computer vision tasks, *e.g.*, scene reconstruction[3, 4, 5, 6], medical identification [7, 8] and motion capture [9, 10].

Many efforts have been made to generate 3D representations from 2D images, which can be roughly divided into two categories. The first category reconstructs the 3D model with multiple views of the same object. With these captured images, classical algorithms usually rely on shading information [11] or Epipolar Geometry constraints, such as Structure from motion [3] and vSLAM [12]. These methods estimate the transformation matrix with the matched keypoints of different images to reconstruct the object. On the other hand, many deep models [13, 14] have been proposed to estimate 3D structure with multi-view inputs. For example, 3D-R2N2 [2] reconstructs and learns the relationship between multi-view images by adopting a 3D recurrent neural network. Kar *et al.* [1] build a 3D grid reasoning model to re-project the multi-view feature to a unified feature grid. Xie *et al.* [15] propose a two-stage coarse-to-fine model to predict the voxel model with a weight-sharing encoder. Methods of this category usually rely on auxiliary information and face difficulties at monocular images.

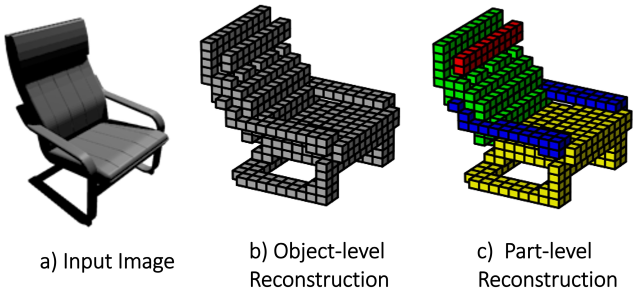

Fig. 1. The motivation of part-level 3D reconstruction. a) Input single-view image. b) Conventional object-level reconstruction task. b) Part-level reconstruction task in this paper.

Methods of the second category reconstruct the 3D model from a single monocular image, which can be more easily applied to unrestricted scenarios. For example,in [16, 17], an Octree-based 3D network is constructed to represent the 3D objects by a convolution and deconvolution architecture. With the proposal of Generative Adversarial network [18], Wu *et al.* [19] leverage the advantages of volumetric convolutional network and generative adversarial network. In [20], a 2.5D sketch is predicted as prior and re-projected to reconstruct the final 3D model. Moreover, there are also some other works to reconstruct 3D models with different representations, such as Point Cloud [21] and mesh model [22]. These methods generate promising results, but these works only focus on the holistic model reconstruction without considering the parts of the model which are crucial for CAD and other local interac-

[*] Jia Li is the corresponding author.

978-1-7281-1331-9/20/$31.00 ©2020 IEEE

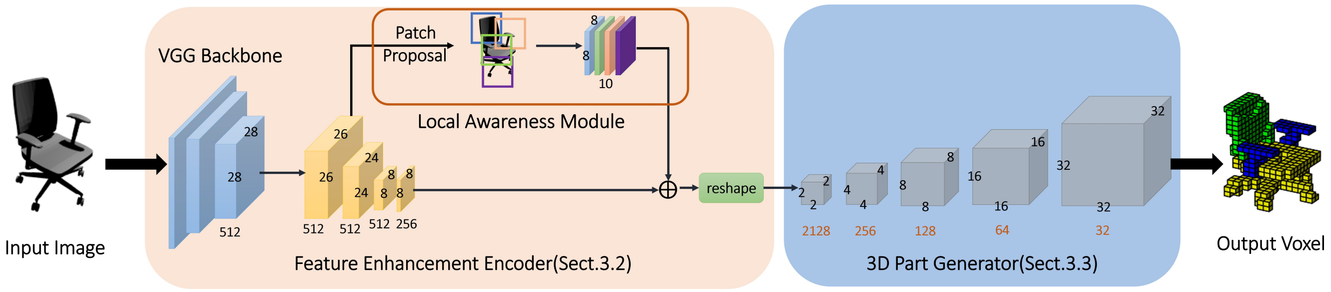

Fig. 2. Overview of our 3D part-level reconstruction framework. Our framework is composed of a feature enhancement encoder and a 3D part generator. The input image is first passed by a VGG backbone to extract the image features and then fed into the feature enhancement module which enhances the local features with a the proposed LAM. After that, the overall features are fed into the 3D part generator to construct the holistic object with part-level information.

tive application usages.

Motivated by the 3D part segmentation model [23] and large-scale part-level annotations [24], it becomes realism to construct the 3D object with fine-grained part structures. As illustrated in Fig. 1, the classical object-level reconstruction task aims to estimate the holistic object from a single monocular input, while the part-level task has the potential to provide fine-grained information, *e.g.*, the armrests and backrests of chairs. Reconstructing objects with part information not only helps the interaction with the object but also the production and animation for man-made CAD models.

To enable the part-level reconstruction from a single-view image, we propose an end-to-end framework to reconstruct the part-level 3D models from a single-view image, which follows the encoder-decoder trends with the local feature enhancement. With the input of a single-view image, we first adopt the typical image encoding network to extract the embedded image-level features. To enhance the local feature representation, we then propose a novel local awareness module to get fine-grained local information from the embedding features. The local features and global image-level features are then fused to form the enhanced features for 3D generation. With the fused image-encoding features, we adopt a 3D part-level generator to reconstruct the holistic object. Our framework can be trained end-to-end with the supervision of the focal part reconstruction loss, which is beneficial to balance the training of parts with different frequencies. To the best of our knowledge, it is the first work to reconstruct the 3D model from a single-view image with part-level information.

Our main contribution can be summarized as: 1) We propose a unified framework for solving the single-view 3D part-level reconstruction problem. 2) We introduce a light-weight local awareness module in the part reconstruction task, which can effectively extract and enhance the local feature representations. 3) Experimental results show that the proposed approach is able to achieve state-of-the-art object-level reconstruction results while simultaneously parsing the part-level information.

2. THE APPROACH

2.1. Overview

In the section, we present the part-level 3D reconstruction framework, which is composed of two modules, *i.e.*, the Feature Enhancement Encoder and 3D Part Generator. Given a single-view input image $\mathcal{I}$, the image feature is first extracted with a common 2D image encoding backbone and then fused with the enhanced feature from the Local Awareness module. With the fused 2D-features together, we rearrange them to a 3D embedding space and finally decoded to a part-level 3D voxel model. The framework is trained end-to-end with the supervision of the part-level voxel annotations $\mathcal{V} = \{\mathcal{V}_i^p, i = 1 \dots C\}$. C denotes the number of parts of each category.

2.2. Feature Enhancement Encoding

To extract discriminative features from single-view images, we resort to the commonly-used 2D Convolutional Neural Network, which serves as backbone encoding in our framework. In this paper, we adopt the VGG-16 backbone network [25], which is pre-trained on the ImageNet dataset to provide high-level information. With the extracted features, we develop a feature enhancement encoding network to enhance the local representations and serve as prerequisites for the 3D generation.

As illustrated in Fig. 2, we first extract the common image-level features with the enhanced encoder (view in yellow). We use 3 × 3 convolutions and stride=2 to build a lightweight feature extractor. The batchnorm [26] and ELU functions are incorporated into different layers. The three encoding features are composed of 512, 512, 256 channels, respectively. To enhance the local representation, we pass the features of the second layer of 512 channels as common feature $\mathbf{F}_c$ and fed into the Local Awareness Module ξ (elaborated in Section 2.3).

With the output feature of local awareness module and

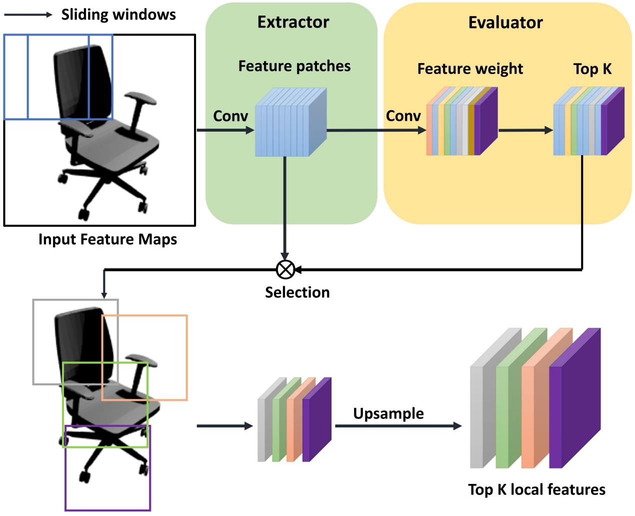

Fig. 3. The illustration of Local Awareness Module. We use sliding window to crop local features to form a patch list. An adaptive feature weight is then calculated to rerank the local features. The top-K discriminative features with high responses are selected to enhance the global feature.

global feature, we then fuse these two outputs with a feature concatenation and 1×1 convolution , which can be formally represented as:

$$\mathbf{F}_{en} = \xi(\mathbf{F}_c) \oplus \delta(W_1(\mathbf{F}_c) + b_1), \tag{1}$$

where W_1 and b_1 are learnable 3×3 convolutional kernels and bias. δ and $\oplus$ denote the ELU activation function and fuse operation, respectively. Then the fused features $(B \times C' \times 8 \times 8)$ are reshaped as $(B \times N \times 2 \times 2 \times 2)$ to form the 3D representation, where B is the batchsize, C' is the corresponding channel size and $N = C' \times 8$.

2.3. Local Awareness Module

In the reconstruction of 3D models, especially in part-level occasions, fine-grained information is in high demand to construct the details of 3D model. To enhance the local feature representation and extract the most useful features, we present an effective Local Awareness module (LAM), which adaptively extract the most useful fine-grained part features to recover the 3D parts.

The LAM is embedded into the feature extractor and fed with the features in deep CNNs. Fig. 3 is a schematic abstraction of LAM, which represents the deep features using images for a better view. The LAM is first encoded with a feature extractor Φ_{enc} and passed by a sliding-window convolution then stack the input features as a patch list. After that these features are convoluted with an evaluator Φ_{eval} to learn the adaptive weighting w_i of ith patch. In this manner, the most discriminative feature for 3D reconstruction will be attached with a higher response, *e.g.*, the patches with rich and characteristic local part details. We thus extract the Top-K local features to form the final output $\xi(\mathbf{F}_c)$, which can be represented as:

$$\xi(\mathbf{F}_c) = \mathcal{H}_k(\Phi_{enc}(\mathbf{F}_c^i); \Phi_{eval}(\mathbf{F}_c^i)), \tag{2}$$

where $\mathcal{H}_k$ denotes the top-K selection function and we set $K = 10$ empirically to balance the local and global features. With the enhanced local selections, the final features are upsampled to be the same size as those of global features. The detailed architecture is exhibited in Tab. 1. The LAM feature encourages the most useful local features in constructing 3D models, while drops the redundant features, *e.g.*, the background region, which are less useful for reconstruction.

Table 1. The architecture of the Local Awareness Module

	Module	Kernel,Stride	Output Shape
Input	-	-	512×26×26
Extractor	Conv	3,1	32×23×23
	BN	-	32×23×23
	ReLU	-	32×23×23
	Conv	3,1	1×23×23
	ReLU	-	1×23×23
	Sliding	4,1	1×(20×20)×4×4
Evaluator	Conv	3,1(pad 1)	32×23×23
	BN	-	32×23×23
	ReLU	-	32×23×23
	Conv	4,1	1×(20×20)
Output	Select K	-	10×4×4
	Bilinear	-	10×8×8

2.4. 3D Part Generator

The 3D part-level reconstruction aims to recover and disentangle different parts in one single process. Unlike the common object-level reconstructions, part-level reconstruction is a more challenging task, especially when the part is extremely small. Keeping this in our mind, we propose the 3D part generator which is adaptively to recover the part-level information.

With the rearranged 3D features from the encoding module, we propose to decode the 3D information from the rich features. The part decoder follows a step-wise operation with 3D deconvolutional kernels, which is illustrated in Fig. 2. We upsample the features by stacking multiple deconvolutional layers and 3D Batch Normalization layers. Finally, the output model is generated with a resolution of 32^3. In the implementation, the kernel size and the stride of each deconvolution layer are set to 4 and 2 respectively, and the numbers of channels are 256, 128, 64, 32 from low resolution to high resolution.

Our training objective is to predict the part label while predicting the over-all shape. Toward this end, the output of

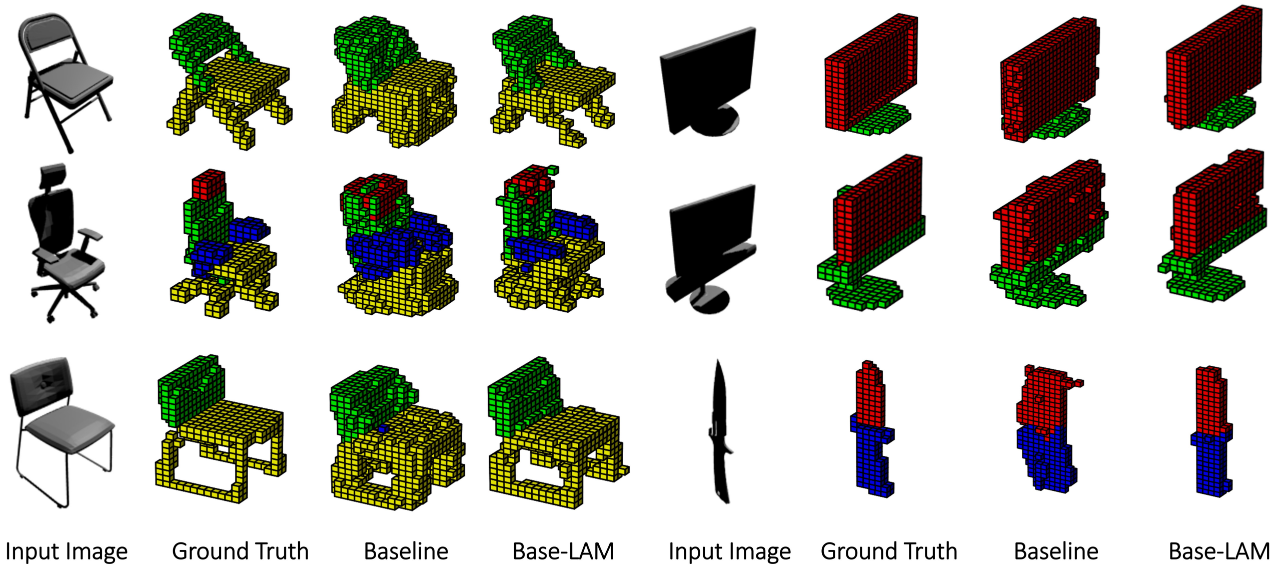

Fig. 4. Visualized reconstruction results of baseline model and our final Base-LAM model.

the generative module is set to contain $\mathcal{C}$ channels on each voxel position $\mathcal{K} = \{k|k = 1, 2..., N_{voxel}\}$, representing the probability value of each part. $\mathcal{C}$ denotes the part number of specific category and N_{voxel} is the number of voxel. It can be trained end-to-end using a multi-class cross-entropy loss:

$$\mathcal{L}_{ce} = -\sum_{k\in\mathcal{K}} \log(p_{tk}), \tag{3}$$

where p_{tk} denote the probability of part label $\mathcal{T} = \{t|t = 0, 1, 2 \ldots \mathcal{C}\}$ at position k.

As discussed above, the distribution of different parts are unbalanced and small parts are usually omitted by large ones. To solve this issue, we develop a 3D focal loss based on [27] as our over-all training objective:

$$\mathcal{L}_{foc} = -\sum_{k\in\mathcal{K}} \alpha_t(1 - p_{tk})^{\gamma} \log(p_{tk}), \tag{4}$$

$$\alpha_t = 1 - \frac{N_t}{\sum_{i\in\mathcal{T}} N_i}, \tag{5}$$

where α_t and γ are hyperparameters to balance the data. We set α_t by inverse class frequency of each category on each voxel in the training set and set $\gamma = 2$ in our experiments.

3. EXPERIMENTS

3.1. Experimental Settings

Dataset. To get the fine-grained part-level 3D annotations, we conduct our experiments based on the large-scale PartNet dataset [24], which consists of 573,585 part instances and 26,671 3D models. We select five representative categories from PartNet and the detailed statistics can be found in Tab. 2. We transform the point cloud annotations to voxel annotations with the label of the corresponding part. For each mesh model, we render 24 images from different horizontal(set cameras every 45° from 0° to 360°) and depression angles (0°, 30°, 60°) and set a constant distance 3 between the model and camera by Blender. The resolution of each image is 224 × 224 and the resolution of each voxel model is 32 × 32 × 32. The training set, validation set, and test set are divided the same as the origin PartNet dataset.

Table 2. Dataset Statistics of PartNet subset in this paper.

Category	Models	part#1	part#2	part#3	part#4	part#5
Chair	6,323	84	6,298	2,875	6,323	-
Table	8,218	65	78	8,039	77	130
Display	928	928	778	-	-	-
Bag	126	126	-	-	-	-
Knife	327	327	74	251	-	-

Implementation details. To train a robust network, we perform data argumentation for the dataset. We firstly randomly change the brightness, contrast, and saturation of the image and add random noise. Secondly, the image randomly left-right flipped and permuted the RGB channels. At last, we normalize the image on RGB channels with the mean and standard deviation calculated from the whole dataset.

We train the model with the learning rate 0.02 and Adam optimizer on a single Nvidia GTX-1080 GPU, and the batch

Table 3. Reconstruction Performance of mIoU on PartNet dataset. Part#i denotes the ith part of each category.

Method	Category	Part#1	Part#2	Part#3	Part#4	Part#5	Object
Base Model	Chair	19.62	34.70	20.69	26.95	-	31.09
	Table	39.58	10.81	25.96	15.92	11.87	25.98
	Display	48.44	34.58	-	-	-	47.99
	Bag	41.68	-	-	-	-	41.68
	Knife	41.39	39.06	43.96	-	-	46.12
Base-LAM	Chair	24.78	46.06	30.80	39.04	-	43.72
	Table	44.72	17.10	32.86	28.79	15.24	32.88
	Display	48.01	38.38	-	-	-	48.04
	Bag	42.70	-	-	-	-	42.70
	Knife	47.88	43.97	51.57	-	-	54.05

Table 4. Comparison on PartNet-Chair Dataset.

Method	mIoU	AP
3D-R2N2 [2]	27.72	42.03
Ours (Base)	31.09	50.85
Ours (Base-LAM)	43.72	62.11

size is set to 8. For each model, we train about 180k iterations and the learning rate decay to 0.002 in the 100k iteration.

Baselines and evaluations. To evaluate our model and the local features module in the dataset, we train a model without the Local Awareness module as our baselines, which follows the same setting as our final model.

In this paper, we choose the mean Intersection over Union (mIoU) and Average Precision (AP) as evaluation criteria for the generation with part label task, and use the mIoU to evaluate the generative quality.

3.2. Comparisons and Evaluations

Part-level reconstruction. We first conduct experiments on the PartNet dataset [24] with 5 representative categories. We first test the performance of our base-model, which is constructed without the Local Awareness enhancement. In Tab. 3, it can be found that the baseline model generates reliable results in constructing the whole object, *e.g.*, 47.99% mIoU of Display. Note that the object-level mIoU is calculated by gathering the disentangle parts as a holistic object. With the local feature enhancement of LAM, our final model improves a large margin on both object and part-level construction. For example, in Tab. 3, the No.3 Part of table category improves from 25.96% to 32.86%, which verifies the effectiveness of our LAM feature enhancement.

The over-all visualized results can be found in Fig. 4. It can be found that our final model (Base-LAM) generate better local details and sharp boundaries, compared to the baseline models. However, compared to the ground-truth label, our generated model still contains some noise voxels, *e.g.*, the armrest of chairs in the second row. This indicates that the 3D part-level reconstruction is a challenging and meaningful task, which motivates further researches and future work.

Object-level reconstruction. To evaluate the performance on object-level reconstruction, we compare our model with the state-of-the-art object reconstruction model 3D-R2N2 [2] in PartNet-Chair subset. In Tab. 4, the results show that our method improves the generative quality. Moreover, it can be seen that our local awareness module is helpful to improve the performance steadily, *e.g.*the AP index improves from 50.85% to 62.11% compared with the base model.

4. CONCLUSIONS

In this paper, we make the first attempt to solve the part-level reconstruction problem from a single-view image, which is a still a less-explored and challenging task. The main problem of the object-level reconstruct method is that these methods miss the fine-grained local features in reconstruction and usually fails to handle the detailed parts. In order to solve this problem, we propose a unified framework with a local feature enhanced representation. We also present an effective and light-weight Local Awareness Module with the selective beneficial local features to promote the reconstruction process. Experimental results show that our model generates reliable part-level structures while achieving state-of-the-art performance in object recovering.

5. ACKNOWLEDGEMENT

This work was partially supported by grants from National Natural Science Foundation of China (No.61922006), and the Beijing Nova Program (Z181100006218063).

6. REFERENCES

[1] Abhishek Kar, Christian Häne, and Jitendra Malik, "Learning a multi-view stereo machine," in *NIPS*, 2017, pp. 365–376.

[2] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction," in *ECCV*, 2016, pp. 628–644.

[3] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer, "A survey of structure from motion*.," *Acta Numerica*, vol. 26, pp. 305–364, 2017.

[4] Greg Slabaugh, Ron Schafer, Tom Malzbender, and Bruce Culbertson, "A survey of methods for volumetric scene reconstruction from photographs," in *Volume Graphics 2001*, 2001, pp. 81–100.

[5] Charles R Dyer, "Volumetric scene reconstruction from multiple views," in *Foundations of image understanding*, pp. 469–489. 2001.

[6] Qian Chen and Gérard Medioni, "A volumetric stereo matching method: Application to image-based modeling," in *CVPR*, 1999, vol. 1, pp. 29–34.

[7] Anastassia Angelopoulou, Alexandra Psarrou, Jose Garcia-Rodriguez, Sergio Orts-Escolano, Jorge Azorin-Lopez, and Kenneth Revett, "3d reconstruction of medical images from slices automatically landmarked with growing neural models," *Neurocomputing*, vol. 150, pp. 16–25, 2015.

[8] Liping Zheng, Guangyao Li, and Jing Sha, "The survey of medical image 3d reconstruction," in *Fifth International Conference on Photonics and Imaging in Biology and Medicine*, 2007, vol. 6534, p. 65342K.

[9] Igor Mordatch, Martin De Lasa, and Aaron Hertzmann, "Robust physics-based locomotion using low-dimensional planning," in *ACM Transactions on Graphics (TOG)*, 2010, vol. 29, p. 71.

[10] Werner Goebl and Caroline Palmer, "Temporal control and hand movement efficiency in skilled music performance," *PloS one*, vol. 8, no. 1, pp. e50901, 2013.

[11] Katsushi Ikeuchi and Berthold KP Horn, "Numerical shape from shading and occluding boundaries," *Artificial intelligence*, vol. 17, no. 1-3, pp. 141–184, 1981.

[12] Hugh Durrant-Whyte and Tim Bailey, "Simultaneous localization and mapping: part i," *IEEE robotics & automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[13] Olivia Wiles and Andrew Zisserman, "Learning to predict 3d surfaces of sculptures from single and multiple views," *International Journal of Computer Vision*, vol. 127, no. 11-12, pp. 1780–1800, 2019.

[14] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang, "Deepmvs: Learning multi-view stereopsis," in *CVPR*, 2018, pp. 2821–2830.

[15] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, Shengping Zhang, and Xiaojun Tong, "Pix2vox: Context-aware 3d reconstruction from single and multi-view images," *arXiv preprint arXiv:1901.11153*, 2019.

[16] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong, "O-cnn: Octree-based convolutional neural networks for 3d shape analysis," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 72, 2017.

[17] Peng-Shuai Wang, Chun-Yu Sun, Yang Liu, and Xin Tong, "Adaptive o-cnn: A patch-based deep representation of 3d shapes," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 217, 2019.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[19] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," in *NIPS*, 2016, pp. 82–90.

[20] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum, "Marrnet: 3d shape reconstruction via 2.5 d sketches," in *NIPS*, 2017, pp. 540–550.

[21] Haoqiang Fan, Hao Su, and Leonidas J Guibas, "A point set generation network for 3d object reconstruction from a single image," in *CVPR*, 2017, pp. 605–613.

[22] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images," in *ECCV*, 2018, pp. 52–67.

[23] Fenggen Yu, Kun Liu, Yan Zhang, Chenyang Zhu, and Kai Xu, "Partnet: A recursive part decomposition network for fine-grained and hierarchical shape segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9491–9500.

[24] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su, "Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding," in *CVPR*, 2019, pp. 909–918.

[25] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.