Pacific Graphics (2021)
M. Okabe, S. Lee, B. Wuensche, and S. Zollmann (Editors)

Short Paper

SM-NET: Reconstructing 3D Structured Mesh Models from Single Real-World Image

Yue Yu, Ying Li, Jing-Yu Zhang, Yue Yang

School of Computer Science & Technology, Beijing Institute of Technology, China

Abstract

Image-based 3D structured model reconstruction enables the network to learn the missing information between the dimensions and understand the structure of the 3D model. In this paper, SM-NET is proposed in order to reconstruct 3D structured mesh model based on single real-world image. First, it considers the model as a sequence of parts and designs a shape autoencoder to autoencode 3D model. Second, the network extracts 2.5D information from the real-world image and maps it to the latent space of the shape autoencoder. Finally, both are connected to complete the reconstruction task. Besides, a more reasonable 3D structured model dataset is built to enhance the effect of reconstruction. The experimental results show that we achieve the reconstruction of 3D structured mesh model based on single real-world image, outperforming other approaches.

CCS Concepts

• ***Computing methodologies*** → ***Reconstruction;*** *Mesh models; Neural networks;*

1. Introduction

Image based 3D model reconstruction is an important subject in the fields of computer graphics, computer vision and machine learning. The biggest difficulty lies in how to complete 3D information missing in 2D. Some approaches use multi-view information and simultaneously solve the projection relationship equation to solve this problem [FP07] [Rob63]. But it is subject to many limitations, such as camera calibration and computational complexity. Therefore, the study of single-view based 3D model reconstruction is widespreadly concerned. Besides, many different 3D shape representations are raised constantly, such as voxel, point cloud, triangular mesh. And with the rapid development of deep neural networks and being widely used in various fields, the research on 3D model reconstruction has also been promoted.

Single-view based 3D reconstruction is of great significance in many fields. For example, it can help 3D modelers perform 3D modeling quickly and easily, and can judge and recognize objects based on images. It plays an important role in many applications, such as industrial manufacturing, intelligent control, virtual reality, etc. Many existing methods of 3D reconstruction only simply approximate the model appearances, but 3D model reconstruction of object is a process of structured understanding. Simply fitting the model cannot adapt to the environment flexibly, and the structural details will be missing during reconstruction. Besides, 3D structured model is easier to modify and design and it is more applicable. Therefore, it is important to infer the structure of object and reconstruct 3D structured model.

In this paper, SM-NET is proposed to reconstruct the 3D structured mesh model from single real-world image. Inspired by PQ-Net [WZX*20], our network regards the entire model as a part sequence and learns to code each part instead of the entire model. It design a part mesh autoencoder to learn to reconstruct model parts and a part sequence autoencoder to place the parts to reconstruct the entire model. First, a shape autoencoder is designed, which encodes various parts of the model through an autoencoder. Further, it encodes the part sequence through a Seq2Seq structure to obtain the latent space of the entire model. Then the decoder decodes the vector from the latent space into a sequence of parts, mapping back to each part and assembling model in order to obtain a subcomponent structured model. Second, we learn the mapping from image to the latent space of autoencoder and further mapping to the model to complete the task of single real-world image based 3D structured mesh model reconstruction.

The input of the proposed network is only a real-world image with complex background. The real-world image has more interference information, and the model reconstruction is more difficult and more challenging. For complex real-world image, the 2.5D generator proposed in MarrNet [WWX*17] is used to extract the 2.5D information in the image, and learn its mapping relationship to the latent space to realize the model reconstruction.

In terms of dataset, the proposal of a large-scale 3D model dataset like ShapeNet [CFG*15] also brings the possibility of deep neural network processing. However, the model used for supervision in this dataset only has the information of the overall shape. Therefore, the method proposed based on ShapeNet can only obtain the overall approximation of the model shape, but cannot re-

DOI: 10.2312/pg.20211388

construct the structural details. Although PQ-Net [WZX*20] uses the PartNet dataset suitable for part splitting in ShapeNet to reconstruct 3D structured model, it is still unreasonable. Based on these experiences, the Part-SM dataset is proposed through improving the PartNet dataset, which reasonably splits the model at an appropriate level and can be well adapted to the task of 3D structured model reconstruction.

Based on this, extensive qualitative and quantitative experiments are conducted on the Part-SM, Pix3D, PASCAL 3D+ datasets to prove that our network can reconstruct higher quality 3D structured mesh models. In summary, our major contributions include: 1) an end-to-end 3D structured model shape autoencoder is proposed; 2) an end-to-end reconstruction from single real-world image to 3D structured mesh model with clear structure is proposed; 3) by improving the Part-Net dataset, a more reasonable Part-SM dataset is established to adapt to the task of 3D structured model reconstruction.

2. Related Work

3D reconstruction based on single real-world image. Modeling the 3D model from 2D image is often limited to the lack of dimensional information, while deep learning method realizes the cross-modal task of model reconstruction by learning the mapping function between the dimensions. MarrNet [WWX*17] and ShapeHD [WZZ*18] used the method of transition information to reconstruct voxel model of the object from the image. Mesh-RCNN [GMJ19] reconstructed the voxel model of the identified object according to the corresponding angle and then meshed it. TM-Net [PHC*19] was based on the sphere deformation method proposed in AtlasNet [GFK*18], in which the topology of mesh model can be changed by designing topology modify network to remove inappropriate patches.

3D structured model reconstruction. Just achieving an approximation of the overall shape of the model can neither allow the network to understand the structure of model, nor restore the details at the structural level. Therefore, how to achieve 3D structured model reconstruction is of great value. GRASS [LXC*17] used a tree structure to recursively obtain the structure of each part of the object, and took the model structure spliced by the 3D bounding box as the output. Im2Struct [NLX18] was based on the GRASS method to decode on the tree structure to realize the generation of single-view 3D structure. A similar work is 3D-PRNN [ZYY*17], which based on depth map generated a series of cuboid shape primitives as parts to spell out the structure of the corresponding model. SAGNet [WWL*19] used a dual-channel variational autoencoder to encode the voxel and the simple structure information of the model, and then it used GRU to fuse the shape features and simple structure features to further realize structured model generation. PQ-Net [WZX*20] used sub-component structure to reconstruct the model, which encoded and decoded the part sequences to reconstruct the structure of the model by a Seq2Seq structure. Our method draws on this idea and achieves better 3D structured mesh model reconstruction.

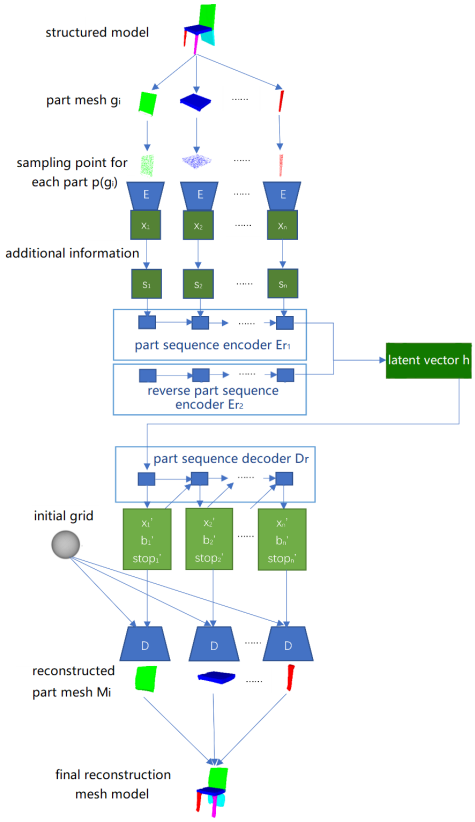

Figure 1: *Network structure of shape autoencoder.*

3. Method

In this section, SM-NET is introduced which learns the mapping from the single image to the latent space of the encoder to realize 3D structured mesh model reconstruction. Given a real-world image, it first trains a network that can autoencode the shape of 3D structured mesh model. Through this process, the network can learn a potential latent space. After that, let the network learn the mapping from images to this latent space, and then the model reconstruction from image is completed.

3.1. Shape Autoencoder Structure

The shape autoencoder is a combination of part mesh autoencoder and part sequence autoencoder. The network structure is depicted in Figure 1. The encoder part of the part mesh autoencoder encodes each part mesh to get a vector sequence. The part sequence encoder encodes this vector sequence into a single vector representing the entire model. Then the part sequence decoder decodes this vector back to a vector sequence, and each vector in it restores the mesh model corresponding to this part through the decoder of the part mesh autoencoder. Finally, these part mesh models are assembled together to obtain an overall 3D structured mesh model. The part mesh autoencoder and part sequence autoencoder are respectively introduced below.

3.1.1. Part Mesh Autoencoder

The part mesh autoencoder is a typical encoder-decoder structure. For each part g_i represented by mesh, record its surface sampling point as $p(g_i)$. Since it is difficult to encode the mesh structure and what we need when decoding is only the shape of the mesh surface, this surface sampling point can replace the mesh for encoding to represent the overall shape of the mesh surface. The PointNet [QSMG17] based encoder E is responsible for encoding $p(g_i)$ into a 1024-dimensional feature vector x_i:

$$x_i = E(p(g_i)) \tag{1}$$

To decode the part mesh from the feature vector x_i representing the surface shape of the part, a template grid deformation method is used. The decoder D design is similar to the deformed network in PointNet [QSMG17], which is a four-layer MLP. The feature vector x_i and the vertex of the initial template grid will be input to the decoder. The decoding process is the process of deforming the mesh by moving the vertex coordinates.

3.1.2. Part Sequence Autoencoder

The part sequence autoencoder is an encoder-decoder structure based on the Seq2Seq structure. Both the encoder and decoder in this structure are implemented using RNN.

The encoder part is a bidirectional stacked RNN [SP97] Er, which is composed of two RNNs Er_1 and Er_2. Each RNN uses GRU as the basic unit. The vector sequence $X = [x_1, x_2, \ldots, x_n]$ obtained by the part mesh encoder represents all parts, where n is the number of parts. Then, the 6-dimensional bounding box information b_i of each part (position and size are each 3-dimensional) and the one-hot vector identifying the total number of parts will be supplemented to obtain the final vector sequence $S = [s_1, s_2, \ldots, s_n]$. This sequence and its reverse sequence $S' = [s_n, s_{n-1}, \ldots, s_1]$ are sent to Er_1 and Er_2 respectively and encoded to h_1 and h_2. The two parts together constitute the model feature vector h:

$$h = [h_1, h_2] = Er(S, S') = [Er_1(S), Er_2(S')] \tag{2}$$

The decoder part is a stacked RNN Dr that can output multiple vectors within each time step. The final feature vector h obtained by the encoder Er is inputed to Dr, and then Dr will output a shape feature x'_i, a bounding box information b'_i and a stop identifier $stop'_i$ at each time step i. The value of the stop identifier $stop'_i$ used to determine whether the sequence should stop is between 0 and 1. When the identifier is greater than 0.5, the number of parts can be considered sufficient, and the decoding iteration process stops:

$$Dr(h) = [x'_1, x'_2, \ldots; b'_1, b'_2, \ldots; stop'_1, stop'_2, \ldots] \tag{3}$$

The shape feature x'_i decoded here will be partially restored to a mesh model by the decoder part of the part mesh autoencoder. At the same time, it is used to determine whether the bounding box information b'_i is used to adjust its position.

3.2. Single Real-world Image Reconstruction Network Structure

To realize the reconstruction from a single real-world image to the 3D model, it is necessary to adjust the network structure for further learning. 2.5D information is data such as normal, depth, contour, etc., which can effectively link 2D and 3D. To map 2D image to 3D model, the 2.5D generator proposed in MarrNet [WWX*17] is used. Baesd on the shape autoencoder in 3.1, the encoding part is replaced with the encoding of the real-world image, and an encoder is constructed to encode 2.5D information, as shown in Figure 2. In this structure, the 2.5D generator obtains a four-channel output O from the real-world image I, and this output O is encoded into a feature vector h' by changing to a four-channel input ResNet-18 [HZRS16] structure Eo:

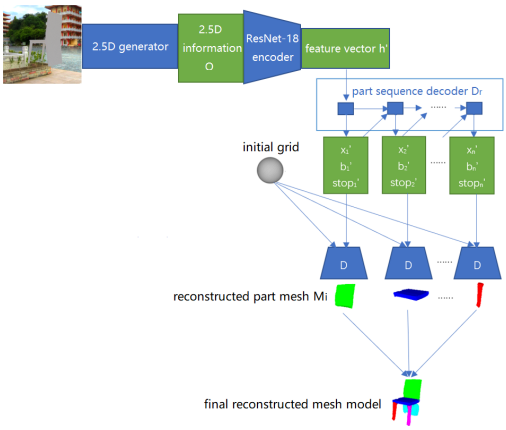

Figure 2: *Network structure for single real-world image reconstruction.*

$$h' = E_o(O) \tag{4}$$

This feature vector h' is then decoded by the part sequence decoder into a vector sequence representing the shape feature of the part. Finally, the part mesh decoder is used to decode the structured mesh model composed of parts through the method of template grid deformation.

3.3. Loss Function

Chamfer distance(CD) loss. Chamfering distance loss is used to supervise the reconstruction of the part mesh models. Given two point sets, the chamfer distance will measure their nearest neighbor distance:

$$\mathcal{L}_{cd} = \sum_{x \in M} \min_{y \in S} \| x - y \|_2^2 + \sum_{y \in S} \min_{x \in M} \| x - y \|_2^2 \tag{5}$$

where $x \in M$ and $y \in S$ are the point sets sampled from the generated mesh surface M and the real mesh surface S respectively.

Shape feature loss. Shape feature loss is used to determine the restoration results of shape features in the part sequence autoencoder. For the shape feature x_i encoded by encoder and the shape feature x'_i decoded by decoder in the part mesh autoencoder, the loss function calculates the mean square error between the two as:

$$\mathcal{L}_x = \frac{1}{n} \sum_{i=1}^{n} \| x'_i - x_i \|_2 \tag{6}$$

Bounding box loss. Similar to shape feature loss, bounding box

loss is used to determine the restoration results of bounding box information. This loss calculates the mean square error between the correct bounding box b_i and the bounding box b_i' decoded by the part sequence decoder as:

$$\mathcal{L}_b = \frac{1}{n}\sum_{i=1}^{n} \parallel b_i' - b_i \parallel_2 \quad (7)$$

Stop loss. Stop loss is used to encourage the decoding process of the part sequence autoencoder to stop in time, so that the number of reconstructed parts is consistent with the input. Stop loss uses binary cross entropy to calculate the difference between $stop_i$ and the correct $stop_i$. At the i-th time step, if the sequence should stop, the value of $stop_i$ is 1, otherwise it is 0:

$$\mathcal{L}_{stop} = \frac{1}{n}\sum_{i=1}^{n}[-stop_i \log stop_i' - (1-stop_i)\log(1-stop_i')] \quad (8)$$

Image coding loss. Image coding loss used for 2.5D information can be correctly mapped to the feature space learned in the shape self-coding process. In order to achieve the mapping from 2.5D information to the latent space correctly, L2 loss is used to supervise the ResNet-18 [HZRS16] encoder:

$$\mathcal{L}_{image} = \parallel h - h' \parallel_2 \quad (9)$$

where h is the actual encoding of model in the latent space, and h' is the output obtained by ResNet-18 encoding the input information.

Each of the above losses will be respectively applied to the appropriate part of the above network, and its weight can be set using hyperparameters.

4. Experiments

4.1. Part-SM Dataset

Like PQ-Net [WZX*20], the PartNet dataset in ShapeNet is used. But there are some problems with PartNet, such as the chair category model. At the finer level, the splitting method is excessive splitting, such as excessively splitting a chair leg into multiple sections. At the coarser level, although the dataset reasonably splits the chair back, chair surface, and armrests, it lacks more specific splits for the chair legs. That means no matter how many legs there are in a model, they will be regarded as the same part. For example, the four legs of a four-legged chair belong to the same component and it is difficult for the network to use a template mesh to reconstruct the four legs in a separated state. Thus, it should be improved and Part-SM dataset is constructed. It keeps the coarsest-grained split of the PartNet dataset except for the chair legs and splits the chair legs by clustering the vertices of the triangle mesh based on a specific threshold τ. After constant adjustment of thresholds and screening, more than 7,000 sets of data without problems and more than 1,000 sets of data with problems in quantity that need to be further processed are obtained.

For 1000 sets of data that need further processing, we find that about 800 groups of four-legged chairs were split into more legs in this batch of data. Therefore, the splitting method of the model is kept which the number of chair legs is 4 in PartNet at the finest level. So the split of this part of the data become reasonable. Finally, more than 8,000 groups of structured models are obtained.

Table 1: *Dataset check result table.*

Category	Quantity	Proportion
Check data	420	100%
Not completely split	18	4.29%
Split error	4	0.95%

In order to check the rationality of the constructed Part-SM dataset, a sampling check was carried out on the dataset, and the results are shown in Table 1. It can be seen that there are still less than 5% of the problem of incomplete separation, mainly because some parts are relatively close, or PartNet separates some separate screws and other structures that are not connected to the chair legs together with the chair legs. Although these parts are not strictly connected, they can be regarded as the same connected part, which has little effect on subsequent work and cannot be regarded as an error. However, the actual proportion of false splitting problems that are indeed to be regarded as errors is very small, less than 1%, and will not have much impact on the whole. So far, the Part-SM dataset is constructed. It improves the PartNet dataset to make it more reasonable.

The experiments are performed on the Part-SM, Pix3D, and PASCAL 3D+ datasets, and demonstrated qualitative and quantitative effects. Pix3D and PASCAL 3D+ are datasets with real-world images and 3D models.

4.2. Training Details

The 2.5D generator was trained with L2 loss for 120 epochs, and the learning rate was 0.001. The part mesh autoencoder was trained for 120 epochs using CD loss. The part sequence autoencoder was trained with shape feature loss, bounding box loss, and stop loss for 2000 epochs on the chair category and 1000 epochs on the lamp category. ResNet, which maps 2.5D information to latent space, was trained for 300 epochs using image coding loss.

4.3. Shape Autoencoder Experiments

To test the quality of the reconstructed structured model in the shape autoencoder, the model reconstructed by SM-Net and PQ-Net [WZX*20] is compared qualitatively and quantitatively on the Part-SM dataset.

Figure 3 shows the results of the qualitative comparison. Different parts in the model are marked with different colors, and the color of the same part in different work should be kept as consistent as possible. As a reconstruction from model to model itself, both of them can get good results on the whole. However, it can be seen that the result of PQ-Net meshing voxels will have obvious voxel traces, and the particles on the surface are obvious. In addition, due to the limitation of voxel resolution in many thin parts, the results obtained by PQ-Net will be thicker than the correct results. This can be observed in the legs of the chair everywhere. The most obvious is in the first row on the right, where PQ-Net has not even reconstructed the connecting rod between the two legs because of its thinness. Further, the second row on the left side can be seen that the PQ-Net generated chair leg voxels intermediate disconnected,

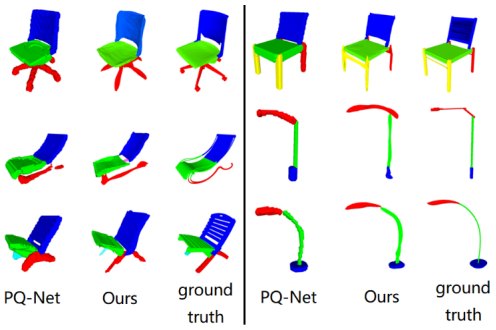

Figure 3: *Visual comparisons for shape autoencoder.*

Table 2: *Quantitative evaluation for shape autoencoder.*

Category	PQ-Net	Ours
Chair	0.00230	**0.00114**
Lamp	0.00537	**0.00249**

which is the possible problems of voxel method. The surface obtained by our method is relatively smooth, and a finer structure can also be obtained.

To evaluate the reconstruction model quantitatively, *CD* will be used as our evaluation indicator. It normalizes the real model and the reconstructed model into a unit cube, then samples the entire surface of the reconstructed model at 10,000 points and calculates the *CD*. Table 2 shows the results of the quantitative comparison. It can be seen that the results we obtained are better than PQ-Net in both categories, which may be due to the granular surface and coarser structure of the voxel.

4.4. Reconstruction Experiments Based on Single Real-world Image

We conducted experiments on the reconstruction of 3D structured mesh model based on single real-world images on the Part-SM, Pix3D, and PASCAL 3D+ datasets. Figure 4 respectively shows the qualitative results on the three datasets. It can be seen that even with background images, we can still generate good structured models and the visual effects are better than others. However, each part only uses a single template mesh to deform, there may still be situations that cannot be fully processed. For example, the void in the middle of the chair legs cannot be obtained in the first row of Figure 4 (a), but the overall structure of the model is still relatively accurate. It is worth mentioning that the void between the armrest and the back of the chair in the second row of Figure 4 (b) still exists after reconstruction. This is the detail between the structures that are easily lost in the integrated model.

Table 3 shows the results of quantitative comparison with CD as the evaluation standard. Like other methods, Pix3D is divided into all set and unobstructed subset to more directly verify the model reconstruction effect. It can be seen that our method is slightly worse than TM-Net [PHC*19] on the Part-SM dataset, but is significantly better than ShapeHD [WZZ*18]. This is because the original TM-Net benefits from the similarity between the training set and the test set. Even if there is no background and other information, TM-Net will also get relatively good results. Besides, the quantitative effects of our method are better than Mesh-RCNN [GMJ19], TM-Net and ShapeHD on Pix3D, and perform well on PASCAL 3D+.

4.5. Applications

Based on the generated 3D structured mesh model, our approach can support the modification and design of the 3D model. This allows us to construct and design the required 3D model more conveniently and quickly. If you want to simply adjust a part in the model, you can remove the corresponding parts in the generated model and process separately. It is disability of the generated model by directly fitting. The model also supports replaceable parts that can be part assembled for different generation models. For example, as shown in Figure 5, you can remove the chair leg parts of the model generated by left image, take the rotating leg part of the model generated by right image and spliced them to the current model. Whether to build a common model directly or to design a unique model that is spliced from different models, our method can achieve them easily and quickly.

5. Conclusion

This paper proposes an end-to-end reconstruction of 3D structured mesh model from single real-world image. It constructs the autoencoder of part mesh and part sequence to get a shape autoencoder, and learns the latent space corresponding to 3D structured mesh model. Later, by learning the mapping from the real-world image to this latent space, the 3D structured mesh model reconstruction from single real-world image is further realized. To achieve the above-mentioned tasks, we also constructed the Part-SM dataset. It is a more reasonable structured mesh model dataset to support our work, and it also has certain application value in some other related work. A large number of qualitative and quantitative results show that our network can complete the reconstruction of the 3D mesh model well and the effect is better.

Acknowledgement

This work is supported by National Natural Science Foundation of China (61807002).

References

[CFG*15] CHANG A. X., FUNKHOUSER T., GUIBAS L., HANRAHAN P., HUANG Q., LI Z., SAVARESE S., SAVVA M., SONG S., SU H., ET AL.: Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015). URL: https://arxiv.org/abs/1512.03012. 1

[FP07] FURUKAWA Y., PONCE J.: Accurate, dense, and robust multi-view stereopsis (pmvs). In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2007), pp. 1–8. doi:10.1109/CVPR.2007.383246. 1

[GFK*18] GROUEIX T., FISHER M., KIM V. G., RUSSELL B. C., AUBRY M.: A papier-mache approach to learning 3d surface generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 216–224. doi:10.1109/CVPR.2018.00030. 2

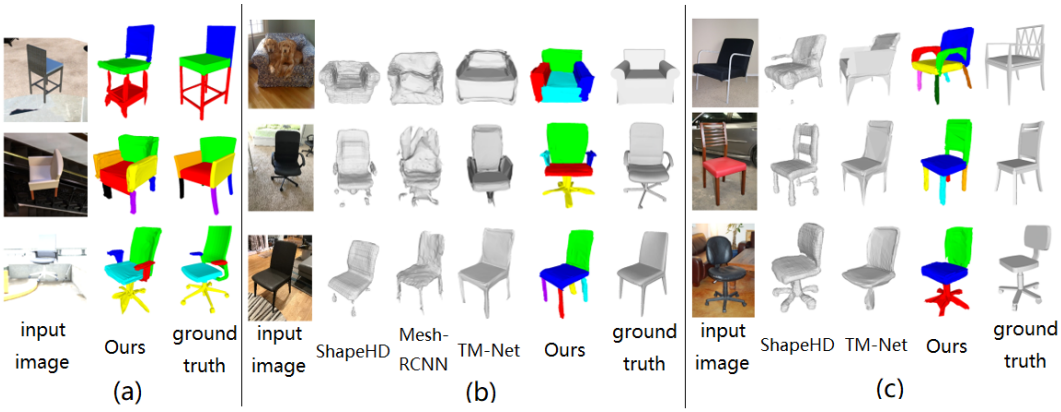

Figure 4: *Qualitative evaluation. (a) Results on Part-SM; (b) Results on Pix3D; (c) Results on PASCAL 3D+.*

Table 3: *Quantitative evaluation.*

Method \ Dataset	Part-SM	Pix3D		PASCAL 3D+
		All set	Unobstructed subset	
Mesh-RCNN	——	0.0258	0.0250	——
ShapeHD	0.01489	0.0176	0.0135	**0.0104**
TM-Net	**0.00859**	0.0207	0.0181	0.0117
Ours	0.01008	**0.0142**	**0.0134**	0.0105

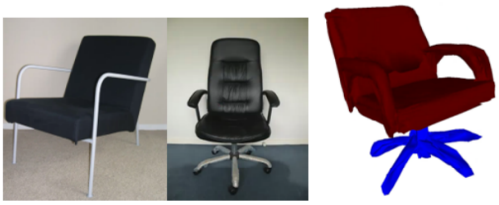

Figure 5: *Part replacement results. Convert the four legs of the left chair into a rotating leg.*

[GMJ19] GKIOXARI G., MALIK J., JOHNSON J.: Mesh r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 9785–9795. doi:10.1109/ICCV.2019.00988. 2, 5

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778. doi:10.1109/CVPR.2016.90. 3, 4

[LXC*17] LI J., XU K., CHAUDHURI S., YUMER E., ZHANG H., GUIBAS L.: Grass: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics (TOG) 36*, 4 (2017), 1–14. doi:10.1145/3072959.3073637. 2

[NLX18] NIU C., LI J., XU K.: Im2struct: Recovering 3d shape structure from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 4521–4529. doi:10.1109/CVPR.2018.00475. 2

[PHC*19] PAN J., HAN X., CHEN W., TANG J., JIA K.: Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 9964–9973. doi:10.1109/ICCV.2019.01006. 2, 5

[QSMG17] QI C. R., SU H., MO K., GUIBAS L. J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 652–660. doi:10.1109/CVPR.2017.16. 3

[Rob63] ROBERTS L. G.: *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963. doi:10.1016/0045-7949(85)90050-1. 1

[SP97] SCHUSTER M., PALIWAL K. K.: Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing 45*, 11 (1997), 2673–2681. doi:10.1109/78.650093. 3

[WWL*19] WU Z., WANG X., LIN D., LISCHINSKI D., COHEN-OR D., HUANG H.: Sagnet: Structure-aware generative network for 3d-shape modeling. *ACM Transactions on Graphics (TOG) 38*, 4 (2019), 1–14. doi:10.1145/3306346.3322956. 2

[WWX*17] WU J., WANG Y., XUE T., SUN X., FREEMAN W. T., TENENBAUM J. B.: Marrnet: 3d shape reconstruction via 2.5 d sketches. *arXiv preprint arXiv:1711.03129* (2017). URL: http://marrnet.csail.mit.edu/. 1, 2, 3

[WZX*20] WU R., ZHUANG Y., XU K., ZHANG H., CHEN B.: Pq-net: A generative part seq2seq network for 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 829–838. doi:10.1109/CVPR42600.2020.00091. 1, 2, 4

[WZZ*18] WU J., ZHANG C., ZHANG X., ZHANG Z., FREEMAN W. T., TENENBAUM J. B.: Learning shape priors for single-view 3d completion and reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 646–662. doi:10.1007/978-3-030-01252-6_40. 2, 5

[ZYY*17] ZOU C., YUMER E., YANG J., CEYLAN D., HOIEM D.: 3d-prnn: Generating shape primitives with recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 900–909. doi:10.1109/ICCV.2017.103. 2