This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Deep Mesh Reconstruction from Single RGB Images via Topology Modification Networks

Junyi Pan[1], Xiaoguang Han[2], Weikai Chen[3], Jiapeng Tang[1], and Kui Jia*[1]

[1]School of Electronic and Information Engineering, South China University of Technology
[2]Shenzhen Research Institute of Big Data, the Chinese University of Hong Kong (Shenzhen)
[3]USC Institute for Creative Technologies

Abstract

Reconstructing the 3D mesh of a general object from a single image is now possible thanks to the latest advances of deep learning technologies. However, due to the nontrivial difficulty of generating a feasible mesh structure, the state-of-the-art approaches [16, 32] often simplify the problem by learning the displacements of a template mesh that deforms it to the target surface. Though reconstructing a 3D shape with complex topology can be achieved by deforming multiple mesh patches, it remains difficult to stitch the results to ensure a high meshing quality. In this paper, we present an end-to-end single-view mesh reconstruction framework that is able to generate high-quality meshes with complex topologies from a single genus-0 template mesh. The key to our approach is a novel progressive shaping framework that alternates between mesh deformation and topology modification. While a deformation network predicts the per-vertex translations that reduce the gap between the reconstructed mesh and the ground truth, a novel topology modification network is employed to prune the error-prone faces, enabling the evolution of topology. By iterating over the two procedures, one can progressively modify the mesh topology while achieving higher reconstruction accuracy. Moreover, a boundary refinement network is designed to refine the boundary conditions to further improve the visual quality of the reconstructed mesh. Extensive experiments demonstrate that our approach outperforms the current state-of-the-art methods both qualitatively and quantitatively, especially for the shapes with complex topologies.

1. Introduction

Image-based 3D reconstruction plays a fundamental role in a variety of tasks in computer vision and computer graphics, such as robot perception, autonomous driving, virtual/augmented reality, *etc.* Conventional approaches mainly leverage the stereo correspondence based on multi-view geometry but are restricted to the coverage provided by the input views. Such requirement renders single-view reconstruction particularly difficult due to the lack of correspondence and large occlusions. With the availability of large-scale 3D shape dataset [3], shape priors can be efficiently encoded in a deep neural network, enabling faithful 3D reconstruction even from a single image. While a variety of 3D representations, e.g. voxels [6, 30, 34] and point cloud [7, 35], have been explored for single-view reconstruction, triangular mesh receives the most attentions as it is more desirable for a wide range of real applications and capable of modeling geometric details.

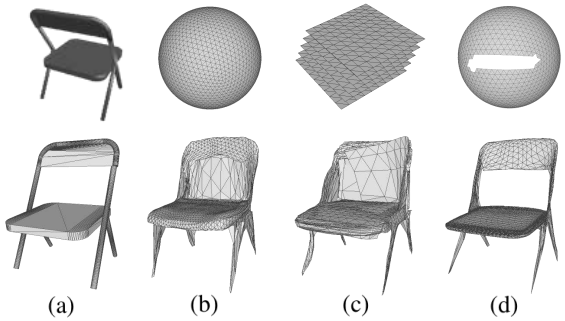

Figure 1. Given a single image of an object (a) as input, the existing mesh-deformation based learning approaches [9] can not well capture the complex topology, regardless of a single (b) or multiple template meshes (c). In contrast, our proposed method is capable of updating the topologies dynamically by removing faces in the initial sphere mesh and achieves better reconstruction results (d).

Recent progresses in single-view mesh reconstruction [32, 9] propose to reconstruct a 3D mesh by deforming a template model based on the perceptual features extracted from the input image. Though promising results have been

*Corresponding author

achieved, the reconstructed results are limited to the identical topological structure with the template model, leading to large reconstruction errors when the target object has a different topology (cf. Figure 1 (b)). Although it is possible to approximate a complex shape with non-disk topology by deforming multiple patches to cover the target surface, there remain several drawbacks that limit its practical usability. Firstly, the reconstructed result is composed of multiple disconnected surface patches, leading to severe self-intersections and overlaps that require tedious efforts to remove the artifacts. Secondly, as obtaining a high-quality global surface parameterization remains a challenging problem, it is nontrivial to generate a proper atlas that can cover the surface with low distortion, only based on a single image. Lastly, it is difficult to determine an appropriate number of surface patches that adapts to varying shapes.

In this work, we strive to generate the 3D mesh with complex topology from a single genus-0 template mesh. Our key idea is a mechanism that dynamically modifies the topology of the template mesh by face pruning, targeting at a trade-off between the deformation flexibility and the output meshing quality. The basic model for deformation learning is a cascaded version of AtlasNet [9] that predicts per-vertex offsets instead of positional coordinates. Starting from an initial mesh M_0, we first apply such deformation network and obtain a coarse output M_1. Then, the key problem is to determine which faces on M_1 to remove. To this end, we propose to train an error-prediction network that estimates the reconstruction error (i.e. distance to the ground truth) of the reconstructed faces on M_1. The faces with large error would be removed to achieve better reconstruction accuracy. However, it remains nontrivial to determine a proper pruning threshold and to guarantee the smoothness of the open boundaries introduced by the face pruning. Towards this end, we propose two strategies to address these issues: 1) a progressive learning framework that alternates between a mesh deformation network, which reduces the reconstruction error, and a topology modification network that prunes the faces with large approximation error; 2) a boundary refinement network that imposes smoothness constraints on the boundary curves, to refine the boundary conditions. Both qualitative and quantitative evaluations demonstrate the superiority of our approach over the existing methods, in terms of both the reconstruction accuracy and the meshing quality. As seen in Figure 1, the proposed method is able to better capture the complex topology with a single sphere template mesh while achieving better meshing quality compared to the state-of-the-art AtlasNet [9].

In summary, our main contributions are:

- The first end-to-end learning framework for single-view object reconstruction that is capable of modeling complex mesh topology from a single genus-0 template mesh.
- A novel topology modification network, which can be integrated into other mesh learning frameworks.
- We demonstrate the advantage of our approach over the state-of-the-arts in terms of both reconstruction accuracy and the meshing quality.

2. Related Works

Reconstructing 3D surfaces from color images has been investigated since the very beginning of the field [27]. To infer 3D structures from 2D images, conventional approaches mainly leverage the stereo correspondences from multi-view geometry [11, 8]. Though high-quality reconstruction can be achieved, stereo based approaches are restricted to the coverage provided by the multiple views and specific appearance models that cannot be generalized to non-lambertian object reconstruction. Hence, learning-based approaches have stood out as the major trend in recent years thanks to its scalability to single or few images.

With the success of deep neural network and the availability of large-scale 3D shape collections, e.g. ShapeNet [3], deep learning-based 3D shape generation has made great progress. In order to replicate the success of 2D convolutional neural network to 3D domain, various forms of 3D representations have been explored. As a natural extension of 2D pixels, volumetric representation has been widely used in recent works on 3D reconstruction [29, 30, 34, 6, 13, 31, 36, 10, 33] due to its simplicity of implementation and compatibility with convolutional neural network. However, deep voxel generators are constrained by its resolution due to the data sparsity and computation cost of 3D convolution. As a flexible form of representing a 3D structure, point cloud has become another major alternative for 3D learning [26, 4] and shape generation [7, 18, 1, 14, 35, 20] due to its high memory efficiency and simple and unified structure. Though enjoying the flexibility to match 3D shape with arbitrary topology, point cloud is not a well suited for imposing geometry constraints, which are critical for ensuring smoothness and appealing visual appearance of the reconstructed surface. Implicit field based 3D reconstruction approaches [21, 22, 5, 23, 13] share the similar advantages with point cloud representation in providing good trade-offs across fidelity, flexibility and compression capabilities. Yet it also remains difficult to regularize the generation of a volumetric implicit field to achieve specific geometry properties.

In contrast, mesh representation is more desirable for real applications since it can model fine shape details and is compatible with various geometry regularizers. Due to the complexity of modifying the mesh topology, most mesh learning approaches strive to obtain a target shape by deforming a template mesh [9, 32, 25, 15, 24] via the learned shape prior. More recently, the advances in differentiable

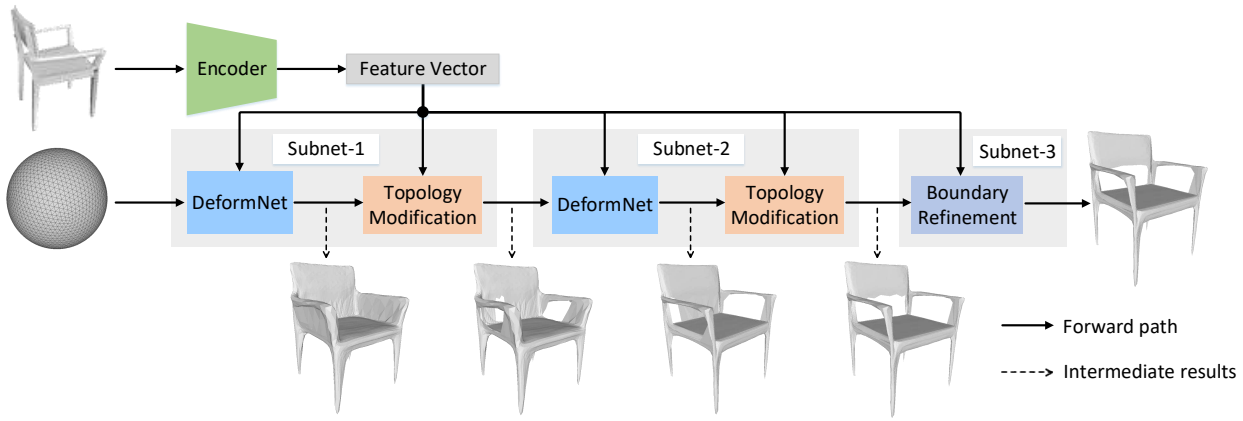

Figure 2. The overview of our pipeline. Given an input image, we first employ multiple mesh deformation and topology modification modules to progressively deform the mesh vertices and update the topologies to approximate the target object surface. A module of boundary refinement is then adopted to refine the boundary conditions.

renderer [19, 16] have proposed to train a mesh generator based on rendering loss, eliminating the need of 3D supervision. However, no prior approaches can dynamically modify the topology of the template mesh, while we propose the first topology modification network that is able to generate meshes with complex topologies from a genus-0 3D model.

3. Topology-adaptive Mesh Reconstruction

Overview. Given a single image I of an object, we attempt to reconstruct the surface S of it. We adopt the triangular mesh as a natural and flexible discretization of the target surface. A mesh is typically defined by $M = (V, E, T)$, where $V \in \mathbb{R}^3$ is the set of mesh vertices, E is the set of edges connecting the neighboring vertices, and T is the set of triangles enclosed by the connected edges. To reconstruct the triangular mesh representation of an object, one could choose to deform a template mesh to approximate the target surface. Nevertheless, the existing deformation-based mesh reconstruction approaches, such as [32, 9, 16], are not allowed to update the faces-to-vertices relationships and thus are restricted by the predefined topology. In order to overcome this limitation, we propose an end-to-end learning pipeline, consisting of three modules, to progressively modify the coordinates and connectivity of the vertices on a predefined mesh M_0. To be specific, the mesh deformation module is adopted to map the vertices on M_0 to the target surface while maintain the connectivity over them; the topology modification module is developed to update the connection relationship between the vertices by pruning the faces which deviate from the ground truth; the boundary refinement module is designed to refine the open boundaries introduced by face pruning. Note that the mesh deformation and topology modification are performed in an alternative manner to gradually recover the overall shape and topology of the target object.

Network structure. We propose a progressive structure to deform a template mesh M_0 to fit the target surface S. In our implementation, M_0 is instantiated as a sphere mesh with 2562 vertices. Figure 2 illustrates the overall pipeline. We leverage an encoder-decoder network for shape generation. On the encoder side, the input image is fed into ResNet-18 [12] to extract a 1024-dimensional feature vector x. The decoder contains three successive subnets. Each of the first two subnets consists of a mesh deformation module and a topology modification module, and the last subnet comprises a single boundary refinement module. Note that each mesh deformation module predicts the per-vertex offset, which can be added to the input mesh to obtain the reconstructed result. The topology modification module then estimates the reconstruction error of the outcome of the preceding deformation module and removes the faces with large error in order to update the mesh topology. Finally, the boundary refinement module enhances the smoothness of the open boundaries to further improve the visual quality.

3.1. Mesh DeformNet

Our mesh deformation module consists of a single multi-layer perceptron (MLP). Specifically, the MLP is composed of four fully-connected layers of size 1024, 512, 256, 128 with non-linear activation ReLU on the first three layers and *tanh* on the final output layer. Given an initial mesh M and the shape feature vector x that contains the prior knowledge of the object, we replicate the vector x and concatenate it with the matrix containing all the vertices of M before feeding them into the MLP. The MLP performs the affine transformation on each vertex of M and generates the vertex displacements. Note that we choose to predict the offsets instead of directly regressing the coordinates. Such a design paradigm enables more accurate learning of fine geometric details with even less training time.

3.2. Topology Modification

To generate objects with various topologies, it is necessary to modify the faces-to-vertices relationship dynamically. Towards this goal, we propose a topology modification network that updates the topological structure of the reconstructed mesh by pruning the faces which deviate significantly from the ground truth. The topology modification network is illustrated in Figure 3, which includes two components: error estimation and face pruning.

3.2.1 Error Estimation

To perform face pruning, it is key to locate the triangle faces that have large reconstruction errors at test time. We propose an error estimation network that predicts the per-face errors of the reconstructed mesh from the preceding mesh deformation network. It leverages a similar architecture with that of the mesh deformation network. In particular, we sample points randomly on the faces of the predicted mesh M and concatenate the replicated shape feature vector x with the matrix containing all the sampling points. The MLP takes as input the feature matrix and predicts the per-point errors (distances to the ground truth). The final error for each triangle face is obtained by averaging the predicted errors for all the sampling points of the triangle face.

3.2.2 Face Pruning

Given the estimated error for each triangle face, we then apply a thresholding strategy that removes the faces whose estimated errors are beyond the predefined threshold to update the mesh topology. However, to obtain a properly pruned mesh structure, the threshold τ needs to be carefully configured: a higher value of τ tends to generate reconstructions with higher errors while a low decision threshold may remove too many triangles and destroy the surface geometry of the generated mesh. To this end, we propose a progressive face pruning strategy that removes error-prone faces in a coarse-to-fine fashion. In particular, we set a higher value for τ at the first subnet and decrease it by a constant factor at the subsequent subnet. Such a strategy enables the face pruning to be performed in a much more accurate manner.

3.3. Boundary Refinement

As shown in Figure 2, a naive pruning of triangles will introduce jagged boundaries that adversely impact the visual appearance. To prevent such artifacts and further improve the visual quality of the reconstructed mesh, we design a boundary refinement module to enhance the smoothness of the open boundaries. It is similar to the mesh deformation module but only predicts the displacement with respect to each input boundary vertex. Note that each boundary vertex is only allowed to move on the 2D plane established by the two boundary edges that intersect at the vertex. We further propose a novel regularization term which penalizes the zigzags by enforcing the boundary curves to stay smooth and consistent. The boundary regularizer is defined as follows:

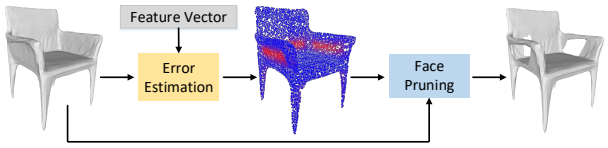

Figure 3. Topology Modification Network. The red color indicates the sampled points with higher estimated errors.

$$\mathcal{L}_{bound} = \sum_{x \in E} \left\| \sum_{p \in \mathcal{N}(x)} \frac{(x-p)}{\|x-p\|} \right\|, \tag{1}$$

where $\{x \in E\}$ is the set of vertices which lie on the open boundary and $\{p \in \mathcal{N}(x)\}$ is the set of neighboring vertices of x on the boundary.

3.4. Training Objectives

Our network is supervised by a hybrid of losses. For mesh deformation and boundary refinement, we employ the commonly-used Chamfer distance (CD) for measuring the discrepancy between the reconstructed result and the ground truth. The error estimation network is trained by the quadratic loss for regressing the reconstruction errors. The boundary regularizer is proposed to guarantee the smoothness of the boundary curves. Besides, we also apply a combination of geometry constraints to regularize the smoothness of the mesh surface during mesh deformation.

CD loss. The CD measures the nearest neighbor distance between two point sets. In our setting, we minimize the two directional distances between the point set randomly sampled from the generated mesh M and the ground truth point set. The CD loss is defined as:

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

where $\{x \in M\}$ and $\{y \in S\}$ are respectively the point sets sampled from the generated mesh M and the ground truth surface S. For each point, CD finds the nearest point in another point set, and sums the squared distances up.

Error estimation loss. We adopt the quadratic loss to train our error estimation network, which is defined as:

$$\mathcal{L}_{error} = \sum_{x \in M} |f_e(x) - e_x|^2, \tag{3}$$

where $\{x \in M\}$ is the point set sampled from the generated mesh M, f_e is the error estimation network, and e_x is the corresponding ground truth error.

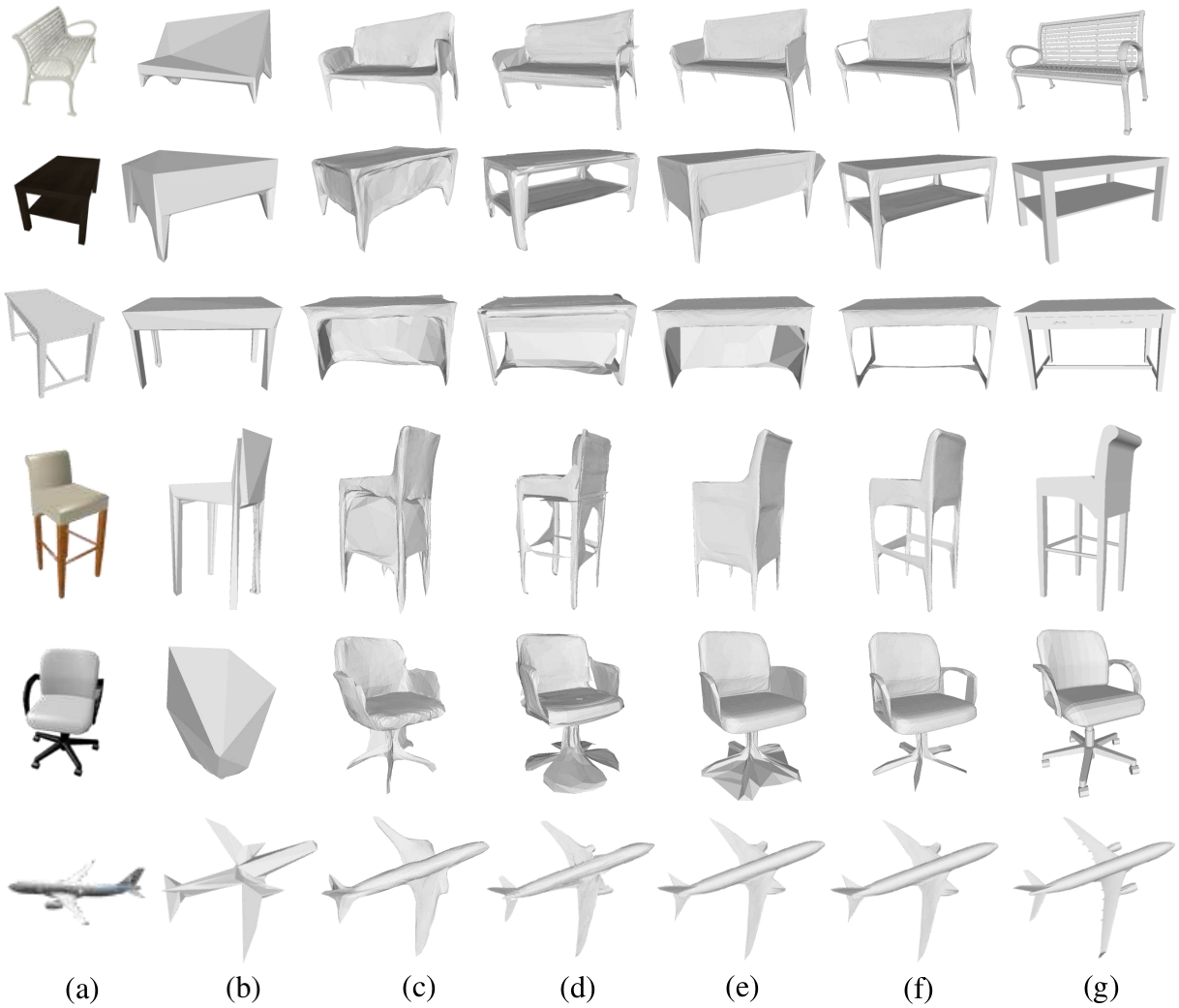

Figure 4. Qualitative results. (a) Input image; (b) N3MR; (c) Pixel2Mesh; (d) AtlasNet-25; (e) Baseline; (f) Ours; (g) Ground truth.

Geometry regularizers. For the mesh deformation module, since the CD loss does not take into account the connectivity of mesh vertices, the predicted mesh could suffer from severe flying vertices and self-intersections. To improve the smoothness of the mesh surface, we add several geometry regularizers. We employ three regularization techniques defined in [32, 16]: the normal loss $\mathcal{L}_{normal}$ which measures the normal consistency between the generated mesh and ground truth, the smoothness loss $\mathcal{L}_{smooth}$ which flattens the intersection angles of the triangle faces and supports the surface smoothness, and the edge loss $\mathcal{L}_{edge}$ which penalizes the flying vertices and overlong edges to guarantee the high quality of recovered 3D geometry.

The final training objective of our system is defined as:

$$\mathcal{L} = \mathcal{L}_{cd} + \lambda_1 \mathcal{L}_{error} + \lambda_2 \mathcal{L}_{bound} + \lambda_3 \mathcal{L}_{normal} + \lambda_4 \mathcal{L}_{smooth} + \lambda_5 \mathcal{L}_{edge}, \quad (4)$$

where λ_i are hyper-parameters weighting the importance of each loss term.

4. Experiments

Dataset. Our experiments are performed on the 3D models collected from five categories in the ShapeNet [3] dataset. To ensure fair comparisons with the existing methods, we adopt the experiment setup in [9]. We use the rendered images provided by [6] as the inputs, where each 3D model corresponds to 24 RGB images. For each 3D shape, $10,000$ points are uniformly sampled on the surface as the ground truth.

Implementation details. The input images all have the same resolution of 224×224. We first train each subnet separately with fixing other components using a batch size of 16 with a learning rate of $1e-3$ (dropped to $1e-4$ after

Category	CD					EMD				
	N3MR	Pixel2Mesh	AtlasNet-25	Baseline	Ours	N3MR	Pixel2Mesh	AtlasNet-25	Baseline	Ours
plane	3.550	2.130	1.566	1.433	**1.390**	10.163	8.859	11.268	8.524	**8.371**
bench	10.865	3.107	2.239	2.950	**2.172**	14.101	10.075	9.808	9.828	**8.713**
chair	15.891	4.787	3.796	4.325	**3.064**	17.246	13.498	11.956	13.313	**10.383**
table	13.438	5.339	4.647	4.798	**3.616**	15.697	11.452	11.562	11.837	**9.604**
firearm	3.230	2.290	1.489	1.145	**1.142**	13.581	8.590	8.711	8.297	**8.226**
mean	9.355	3.531	2.747	2.930	**2.277**	14.158	10.495	10.661	10.360	**9.059**

Table 1. Quantitative comparison with the state-of-the-art methods. The CD and EMD are computed on $10,000$ points sampled from the generated mesh after performing ICP alignment with the ground truth. The CD is in units of 10^{-3} and the EMD is in units of 10^{-2}.

200 epochs) for 300 epochs. The entire network is then fine-tuned in an end-to-end manner. The values of hyper-parameters used in Equation (4) are $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 1e-2$, $\lambda_4 = 2e-7$, $\lambda_5 = 0.1$. The threshold τ for face pruning is set to be 0.1 at the first subnet and decreased by a factor of 2 at the subsequent subnet.

4.1. Comparisons with the State-of-the-arts

We first compare the performance of our approach with three state-of-the-art methods for single view 3D mesh reconstruction, including Neural 3D Mesh Renderer [16] (N3MR), Pixel2Mesh [32] and AtlasNet [9] with 25 patches (AtlasNet-25). We also compare with the baseline approach which refers to our framework without the topology modification and boundary refinement module.

Qualitative comparisons. The visual comparison results are shown in Figure 4. While N3MR can reconstruct the rough shapes, it fails to capture the fine details of the geometry and is not able to model surface with non-disk topology. Pixel2Mesh performs generally better than N3MR in terms of the capability of modeling the fine structures. However, as Pixel2Mesh employs a similar mesh deformation strategy, it struggles to reconstruct shapes with complex topologies, especially for the chairs and tables. Our baseline approach also has the same problem as the topology modification module is not applied. Thanks to the use of multiple squares as the template model, AtlasNet-25 can generate meshes with various topologies. However, it suffers from severe self-intersections and overlapping and still fails to reconstruct some instances with more complex topologies, e,g, the desk in the third row and the chair in the fifth row. In comparison, our approach has outperformed the other approaches in terms of visual quality. We are able to generate meshes with complex topologies while maintaining high reconstruction accuracy thanks to the topology-modification modules. In addition, our method scales well to the shapes with simple topologies. For the objects that can be well reconstructed from a template sphere (e.g. the plane), the spherical topology is faithfully preserved.

Method	CD	EMD
AtlasNet-25 (PSR)	3.430	12.574
Ours (PSR)	**2.304**	**10.415**

Table 2. Quantitative comparison with AtlasNet-25 after PSR.

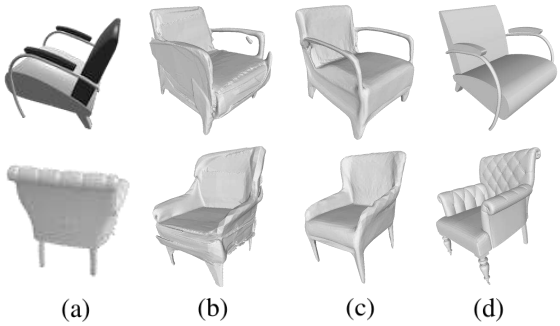

Figure 5. Qualitative comparison with AtlasNet-25 after PSR. (a): Input images; (b): AtlasNet-25; (c): Ours; (d): Ground truth.

Quantitative comparisons. We adopt the widely used Chamfer Distance (CD) and Earth Mover's Distance (EMD) to quantitatively evaluate the results. Both metrics are computed between the ground truth point cloud and $10,000$ points uniformly sampled from the generated mesh. Since the outputs of Pixel2Mesh [32] are non-canonical, we align their predictions to the canonical ground truth by using the pose metadata available in the dataset. Additionally, we apply the iterative closest point algorithm (ICP) [2] on all the results for finer alignment with the ground truth. The quantitative comparison results are shown in Table 1. Our approach consistently outperforms the state-of-the-art methods in both metrics over all five categories, especially on the models with complex topologies (e.g. chair and table).

Poisson surface reconstruction. Although our method can generate visually appealing meshes with smooth surfaces and complex topologies, it still has the inherent draw-

Figure 6. Qualitative comparison with AtlasNet-25 on real images. (a): Input images; (b): Input masks; (c): AtlasNet-25; (d): Ours.

back of producing open surfaces due to the face pruning operations. To avoid the above mentioned drawbacks, one could densely sample the surface and reconstruct the mesh from the obtained point cloud. In particular, we first sample $100,000$ points, together with their oriented normals, from the reconstructed surface and then apply Poisson surface reconstruction (PSR) [17] to produce a closed triangle mesh. To evaluate the performance of applying PSR on our results, we quantitatively compare with AtlasNet-25. Specifically, for both methods, we randomly selected 20 shapes from the chair category, and run the PSR algorithm to get the corresponding closed meshes. In Table 2, we show the quantitative comparisons measured in CD and EMD. As seen from the results, our approach generates more accurate results under both measurements. We show the visual comparisons in Figure 5. Note that we generate meshes with significantly fewer artifacts and correct topologies compared to the AtlasNet, proving the better meshing quality of our method.

Reconstructing real-world objects. To qualitatively evaluate the generalization performance of our method on the real images, we test our network on the Pix3D [28] dataset by using the model trained on the ShapeNet [3]. Figure 6 shows the results reconstructed by our method and AtlasNet, where the objects in the images are manually segmented. Our approach is still able to faithfully reconstruct a variety of objects with complex topologies and achieves better accuracy compared against AtlasNet, indicating that our method scales reasonably well on the real images.

4.2. Ablation Studies

Robustness to initial meshes. We first test our approach with different initial meshes (e.g. sphere and unit square). As seen from the visualization results in Figure 7, our method achieves similar performance with the two different initial meshes, demonstrating the robustness of our method.

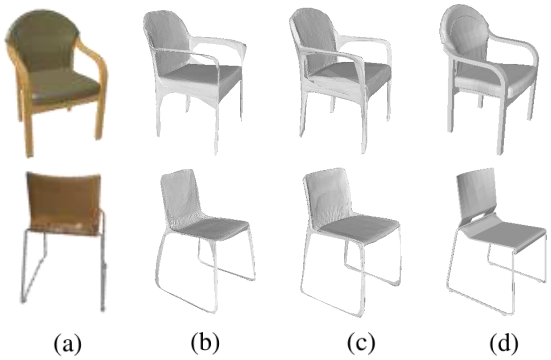

Figure 7. Qualitative results with different initial meshes. (a): Input images; (b): Unit square; (c): Sphere; (d): Ground truth.

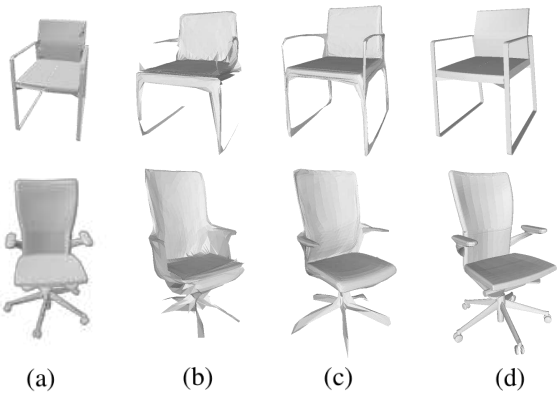

Figure 8. Ablation study on progressive shaping. (a): Input images; (b): Reconstructions w/o progressive shaping; (c): Reconstructions with progressive shaping; (d): Ground truth.

Progressive shaping. Our proposed architecture consists of multiple mesh deformation and topology modification modules that progressively recover the 3D shape. To validate the effectiveness of such progressive shaping strategy, we retrain our network with removing the first subnet in the decoder. Figure 8 shows the visualization results. Without progressive shaping, the face pruning cannot be performed in an accurate manner, which could destroy the surface geometry of the generated mesh.

Face pruning threshold. We investigate the effect of the threshold τ by using 20% of training samples as the validation set, where Chamfer distance (CD) is used as the measure that sums two directional reconstruction errors of $\text{Prediction} \rightarrow \text{GT}$ and $\text{GT} \rightarrow \text{Prediction}$ (cf. Equation (2)). Figure 9 plots the results, suggesting that $\tau \in [0.05, 0.2]$ strikes a good balance between the two directional distances. We thus set $\tau = 0.1$ in all our experiments.

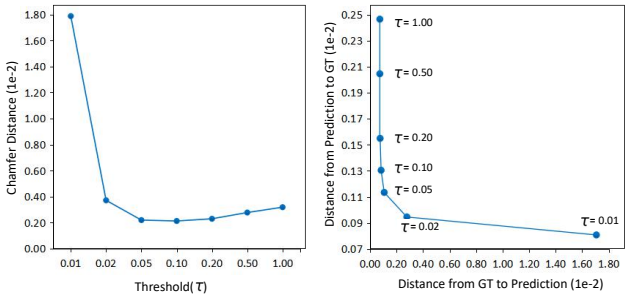

Figure 9. Effect of τ values on Chamfer distance, distance from GT to prediction, and distance from prediction to GT.

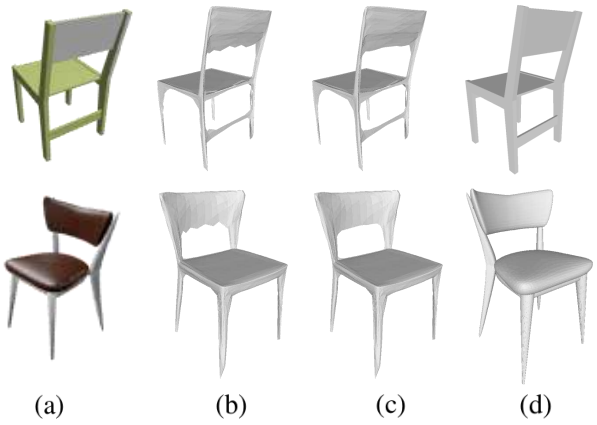

Figure 10. Ablation study on boundary refinement. (a): Input images; (b): Reconstructions w/o boundary refinement; (c): Reconstructions with boundary refinement; (d): Ground truth.

Boundary refinement. To demonstrate the effectiveness of the proposed boundary refinement module, we show the visualized reconstruction results with and without the boundary refinement in Figure 10. By using the proposed boundary refinement module, one can achieve significantly cleaner mesh with higher visual quality.

4.3. Shape Autoencoding

Besides the single-view 3D reconstruction, our framework can also be applied for 3D shape autoencoding. In this section, we demonstrate the capability of our approach to reconstruct meshes from the input point clouds. Toward this goal, we randomly select $2,500$ points from the ground-truth point cloud and employ the PointNet [26] to extract the corresponding latent features. Again, we compare both the quantitative and qualitative results against the state-of-the-art AtlasNet [9]. To ensure fair comparisons, we use the same experiment settings in [9]. The results are shown in Table 3 and Figure 11. As shown in the results, our approach achieves superior performance both qualitatively and quantitatively.

Method	CD	EMD
AtlasNet-25 (AE)	0.765	8.467
Ours (AE)	**0.655**	**6.754**

Table 3. Quantitative results of 3D shape autoencoding. The results take the means on the five shape categories used in the single-view reconstruction.

Figure 11. Qualitative results of 3D shape autoencoding. (a): Ground truth Meshes; (b): AtlasNet-25; (c): Ours.

5. Conclusion

We have proposed an end-to-end learning framework that is capable of reconstructing meshes of various topologies from single-view images. The overall framework includes multiple mesh deformation and topology modification modules that progressively recover the 3D shape, and a boundary refinement module that refines the boundary conditions. Extensive experiments show that our method significantly outperforms the existing methods, both quantitatively and qualitatively. One limitation of our method is the inherent drawback of producing non-closed meshes. But it can be resolved by a post-processing procedure that reconstructs closed surfaces from densely sampled point clouds. Future research directions include designing a differentiable mesh stitching operation to stitch the open boundaries introduced by the face pruning operations.

6. Acknowledge

This work is supported in part by the National Natural Science Foundation of China (Grant No.: 61771201), the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (Grant No.: 2017ZT07X183), the Pearl River Talent Recruitment Program Innovative and Entrepreneurial Teams in 2017 (Grant No.: 2017ZT07X152), the Shenzhen Fundamental Research Fund (Grants No.: KQTD2015033114415450 and ZDSYS201707251409055), and Department of Science and Technology of Guangdong Province Fund (2018B030338001).

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International Conference on Machine Learning*, pages 40–49, 2018.

[2] PJ Besl and Neil D McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[4] Weikai Chen, Xiaoguang Han, Guanbin Li, Chao Chen, Jun Xing, Yajie Zhao, and Hao Li. Deep rbfnet: Point cloud feature learning using radial basis functions. *arXiv preprint arXiv:1812.04302*, 2018.

[5] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 628–644. Springer, 2016.

[7] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2463–2471, 2017.

[8] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2010.

[9] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018.

[10] Xiaoguang Han, Zhen Li, Haibin Huang, Evangelos Kalogerakis, and Yizhou Yu. High-resolution shape completion using deep neural networks for global structure and local geometry inference. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 85–93, 2017.

[11] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[12] 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*, pages 770–778, 2016.

[13] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 336–354, 2018.

[14] Eldar Insafutdinov and Alexey Dosovitskiy. Unsupervised learning of shape and pose with differentiable point clouds. In *Advances in Neural Information Processing Systems*, pages 2802–2812, 2018.

[15] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018.

[16] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2018.

[17] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):29, 2013.

[18] Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[19] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

[20] Priyanka Mandikal, K L Navaneet, Mayank Agarwal, and R Venkatesh Babu. 3D-LMNet: Latent embedding matching for accurate and diverse 3d point cloud reconstruction from a single image. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[21] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[22] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Deep level sets: Implicit surface representations for 3d shape inference. *arXiv preprint arXiv:1901.06802*, 2019.

[23] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4480–4490, 2019.

[24] Junyi Pan, Jun Li, Xiaoguang Han, and Kui Jia. Residual meshnet: Learning to deform meshes for single-view 3d reconstruction. In *International Conference on 3D Vision (3DV)*, pages 719–727, 2018.

[25] Jhony K Pontes, Chen Kong, Sridha Sridharan, Simon Lucey, Anders Eriksson, and Clinton Fookes. Image2mesh: A learning framework for single image 3d reconstruction. In *Asian Conference on Computer Vision*, pages 365–381, 2018.

[26] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[27] Lawrence G Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963.

[28] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2974–2983, 2018.

[29] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2088–2096, 2017.

[30] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2634, 2017.

[31] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018.

[32] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018.

[33] Shuco Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)*, 37(4):162, 2018.

[34] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, pages 1696–1704, 2016.

[35] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 206–215, 2018.

[36] Yi Zhou, Liwen Hu, Jun Xing, Weikai Chen, Han-Wei Kung, Xin Tong, and Hao Li. Hairnet: Single-view hair reconstruction using convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018.