IMMAT: Mesh Reconstruction from Single View Images by Medial Axis Transform Prediction

Jianwei Hu, Gang Chen

BNRist and School of Software, Tsinghua University, China

Baorong Yang

College of Computer Engineering, Jimei University, China

Ningna Wang, Xiaohu Guo

Department of Computer Science, The University of Texas at Dallas, USA

Bin Wang*

BNRist and School of Software, Tsinghua University, China

Abstract

The representation of a 3D shape is a key element for capturing the overall structure as well as the local details. In this paper, we propose to predict a mesh representation of the Medial Axis Transform (called medial mesh) as an intermediate representation with our *IMMAT* framework, to reconstruct the 3D shape from a single view image. Because the MAT contains the skeleton topology and local thickness information, it not only enhances the ability to reconstruct topologically complex shapes but also better preserves the local details with its thickness control. The framework consists of three modules. The *Image2Sphere* module predicts the medial spheres inside the shape surface and the *Topology Prediction* module predicts the topological relationship (skeleton) between the predicted spheres. Then the *MAT Smoothing* module smooths the predicted MAT and fine-tunes the coordinates and radii of the spheres by graph convolution. The final 3D surface can be reconstructed by converting the predicted MAT to an implicit surface through CSG operation and then extracting the boundary surface through Marching Cubes. Experimental results show that our method outperforms the state-of-the-art methods both quantitatively and qualitatively on the reconstruction task.

Keywords: Deep Learning, Medial Axis Transform, 3D Reconstruction, Single View Image.

1. Introduction

Inferring a 3D shape from a single view image has received much attention in recent years but is still a very challenging problem in various tasks of computer vision and computer graphics. With the availability of large-scale 3D shape datasets, such as ShapeNet [1], deep learning based approaches can generate 3D shapes with representations of volume [2, 3, 4, 5, 6], point clouds [7, 8], or triangular mesh [9] as the output of neural networks.

Geometry and topology are two important features of a 3D shape and shapes are often visually different from each other due to the difference in geometry and topology. Point clouds and voxels only express the geometry and have poor ability to learn the topology of 3D shapes.

Triangular mesh expresses geometry and topology at the same time. However, it is difficult to learn surface mesh from a single view image by convolutional neural networks. The methods based on template mesh deformation [9, 10] have achieved promising results, but they can only reconstruct shapes of very limited topologies that are often not complex enough. Eliminating invalid triangular faces which cause the incorrect topology can break through the topological constraint of given templates, but it will destroy the closure of a mesh and cause boundary distortion.

The skeleton-based method [10] has been proposed to capture the underlying topological structure of the target object. It is effective for reconstructing topologically complex shapes. However, the predicted skeleton points only provide an initial topology, which lacks geometric information to directly reconstruct the surface mesh. To learn better geometric structures, the skeleton points need to be transferred into voxels and meshes. This transfer inherits the disadvantage of mesh deformation, which may lead to self-intersection of the mesh or even destroy the initial topology. The whole pipeline does not consider the thickness of the shape and leads to an uneven surface in the generated mesh that seriously affects the visual effect.

Inspired by the skeleton-based method, we propose to construct the Medial Axis Transform (MAT) [11] of a 3D shape from a single view image. Different from skeleton points which are point clouds on the skeleton, MAT has more outstanding characteristics:

1) MAT uses medial spheres located on the skeleton with radii to represent the geometry. The radius is the distance from the

*Corresponding author

Email addresses: hjw17@tsinghua.mails.edu.cn (Jianwei Hu), g-chen21@mails.tsinghua.edu.cn (Gang Chen), yangbaorong@jmu.edu.cn (Baorong Yang), nxw180011@utdallas.edu (Ningna Wang), xguo@utdallas.edu (Xiaohu Guo), wangbins@tsinghua.edu.cn (Bin Wang*)

center of the sphere on the skeleton to the surface of the shape, which represents the local thickness and can be used for surface reconstruction.

2) MAT has connection relationships among medial spheres to represent the topology information of the shape. The connections represent the skeleton structures and can flexibly reconstruct various complex shapes. An edge between two spheres expresses the curve structure, and a face among three spheres expresses the surface structure.

3) A MAT can directly recover a manifold and watertight triangular mesh by Marching Cubes [12]. Therefore, only the MAT representation is operated throughout the whole pipeline, without the need to transfer to voxel and mesh representations like the skeleton-based method [10].

In this paper, we propose *IMMAT* to predict MAT to directly learn the medial spheres and skeleton topology of a 3D shape from a single view image. Different from the Point2Skeleton [13] which learns a MAT from point clouds (the input and output are in the same 3D space), our task to solve the gap between 2D and 3D is more challenging. In our framework, we divide the MAT prediction into three stages and propose the corresponding deep network modules. Firstly, the Image2Sphere module predicts a set of discrete spheres with different radii from a single view image. Then the Topology Prediction module predicts the topological relationships between these spheres to construct the topology of MAT. We further use the MAT Smoothing module to smooth the spheres of MAT and improve the quality of the reconstructed surface mesh. Fig. 1 shows an overview and several basic shapes from different geometries and topologies. We will release the code and MAT datasets to the public for further research. The main contributions of this paper include:

- We introduce MAT as the underlying representation for shape reconstruction from a single view image and propose a novel framework for MAT prediction. We have created a MAT dataset that will be open source and used for deep learning research.
- We propose the Image2Sphere module, the first learning-based method for predicting medial spheres from a single view image, to simultaneously predict the spatial distribution and volume information of 3D shapes.
- We propose a deep learning based method to predict the topology relationships of 3D spheres and achieve high-quality reconstruction results with the generated MAT.

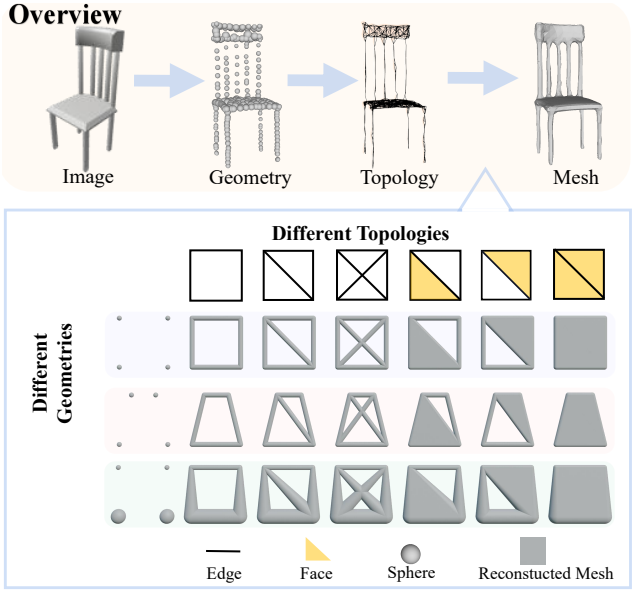

Figure 1: Our proposed approach can generate a closed watertight surface mesh from a single view RGB image, by precisely predicting the geometry (medial spheres) and complex topology (edges and faces) with a MAT representation. At the bottom box, we show some examples of reconstructed meshes in different geometries and typologies of MAT, which reflects our ability to generate complex shapes. The first row shows six different typologies of four spheres with corresponding geometries (sphere centers and radii) in the second row. The following two rows indicate the impact of changing geometries by updating the locations of sphere centers or sizes of radii. The meshes at each column have the same topology but different geometries.

2. Related Work

Mesh-based deformation methods learn the vertices' positions and deform an initial mesh (*e.g.*, an ellipsoid) to achieve similarity in the overall shape [14, 15]. But it is not capable of generating shapes of arbitrary topology from a genus-0 mesh. Deformation from a similar template mesh [16] further enhances similarity in overall shape and local details. But because it does not change the topology of the source template, these methods can only reconstruct surfaces with fixed topology. Topology modification method [17] prunes the redundant edges/faces of the triangle mesh, enabling the evolution of topology and improving the local details. However, the rough pruning operations could potentially destroy the watertight property of the generated mesh. PSG [8] generates point clouds from a single view image.

The skeleton-based method [10] splits the shape reconstruction task into three stages. First, some meso-skeleton points are predicted and converted into a volumetric representation. After refinement, a base mesh similar to the target is extracted. Finally, a mesh deformation network is used to produce geometric details.

In addition to explicit representations, implicit representations have become popular in recent studies. Occupancy Network [18] learns a continuous occupancy function as the representation of a 3D shape with neural networks. DeepSDF [19], DISN [20] predict signed distance functions of 3D points near the 3D surface. SIFs [21] represents a 3D shape by combining a set of shape elements (structured implicit functions). The element is a scaled axis-aligned anisotropic 3D Gaussian, and the whole 3D shape is represented as the sum of these shape elements. DSIFs [22] provides local geometry details by adding deep neural networks as deep implicit functions (DIFs). LDIF [23] performs well on local shape details of 3D reconstruction.

There are some recent works on exploiting MAT [24] as an underlying representation for shape analysis. MAT-Net [25] validates the performance of MAT representation in the 3D shape classification task. P2MAT-NET [26] learns the pattern of sparse point clouds and transforms them into spheres and then reconstructs the connectivity of spheres with a post-processing manner to approximate MAT. Point2Skeleton [13] proposes an unsupervised method to learn the MAT representation from point clouds, which can be used for shape reconstruction or segmentation of point clouds.

3. The Method

The overall goal of this work is to reconstruct a surface **O** from an image **I** of a single object by predicting MAT of the shape from **I**.

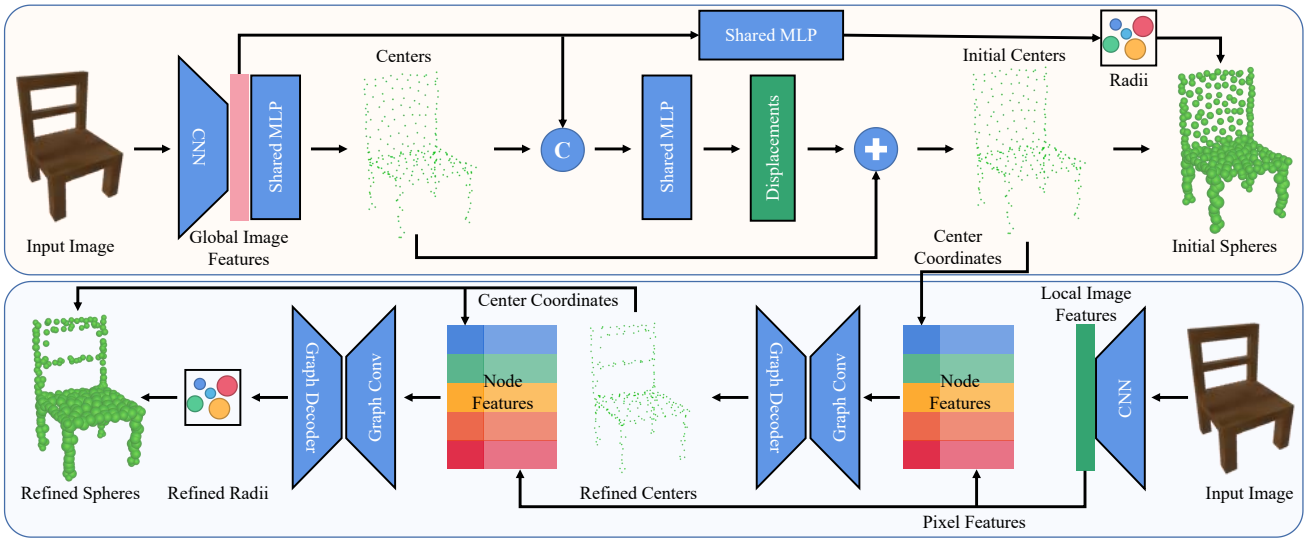

Figure 2: The overall pipeline of the Image2Sphere module.

We follow the definition of MAT in Q-MAT [27]. The MAT $\mathbf{M}$ of a 3D shape is composed of spheres $\mathbf{S}$, edges $\mathbf{E}$, and faces $\mathbf{F}$, as shown in Fig. 1. We define $\mathbf{M} = (\mathbf{S}, \mathbf{E}, \mathbf{F})$, each sphere $\mathbf{s}$ is denoted as a 4D vector $\mathbf{s} = (\mathbf{c}, r)$ with the center $\mathbf{c}$ and radius r of the sphere. $\mathbf{e}_{ij} = \{i, j\}$ is the edge defined by linear interpolation of two medial spheres $(1-t)\mathbf{s}_i + t\mathbf{s}_j$, $t \in [0, 1]$. Similarly, a medial face (also called medial slab), $\mathbf{f}_{ijk} = \{i, j, k\}$, is a convex combination of three medial spheres $a_1\mathbf{s}_i + a_2\mathbf{s}_j + a_3\mathbf{s}_k$ with $a_i \geq 0$ and $a_1 + a_2 + a_3 = 1$. MAT preserves the topology and volume information of the object and can be represented with any resolution (number of spheres), which balances the complexity and the fineness of the reconstructed mesh.

Our method consists of three modules: Image2Sphere, Topology Prediction, and MAT Smoothing. In Image2Sphere, we learn the initial spheres from the global feature of an input image, which achieves certain similarities in appearance. Then we use local image features to refine the coordinates of the initial spheres and predict their new radii. In Topology Prediction, two local adjacency matrices are predicted from N refined spheres and their $N \times K$ neighbor spheres. For each sphere $\mathbf{s}_i$, we first use K-Nearest Neighbors (K-NN) to query K neighbor spheres, then a convolutional neural network is trained to obtain local features from K spheres, and finally the fully connected layers are used to predict an edge probability matrix $\mathbf{ME}_i$ and a face probability matrix $\mathbf{MF}_i$. $\mathbf{ME}_i$ is a $1 \times K$ vector that denotes the probability of edges between $\mathbf{s}_i$ and its K neighbors. $\mathbf{MF}_i$ is a $1 \times K \times K$ matrix, with each entry $\mathbf{MF}_{i,j,k}$ representing the probability of face between $\mathbf{s}_i$ and its two neighbors $\mathbf{s}_j$ and $\mathbf{s}_k$. Finally, the edges and faces with higher probability together with the predicted spheres from the Image2Sphere form a predicted MAT. Ideally, the connected spheres should have similar coordinates and radii distribution. With the predicted spheres and topologies, the MAT smoothing module finetunes the spheres' centers and radii to smooth the surface and curve structures.

3.1. Image2Sphere

The Image2Sphere module is proposed to predict a precise distribution of spheres from the input color image. Note that the sphere centers are located on the skeleton, not on the surface. It includes two sub-networks: generating initial spheres using global image features and generating refined spheres using local image features, as shown in Fig. 2. We firstly use ResNet18 to encode the image into a global feature vector, then decode it into centers of spheres with multi-layer perceptrons (MLPs). To get more accurate sphere predictions, a small displacement is applied to the centers. This displacement is decoded using MLPs by concatenating the global image features and the centers. For radius prediction, the global image features are also decoded into an $N \times 1$ vector that contains the radius of each sphere. As a result of this stage, initial spheres, including the initial centers and the initial radii, are predicted and can be used to reconstruct a simple shape. We firstly use ResNet18 to encode the image into a global feature vector, then decode it into a set of spheres. The centers and radii of spheres are decoder by two multi-layer perceptrons (MLPs) respectively. However, the initial spheres might not be able to capture the fine details of the shape. For example, as shown in the top part of Fig. 2, the predicted initial spheres could not fit well at the back of the chair. Therefore, local features extracted from the input image are introduced to optimize the predicted spheres. Similar to Pixel2Mesh [14], we use camera parameters of the image to project the centers of initial spheres onto the 2D image and extract the corresponding pixel features from the image feature maps. Then we combine pixel features with centers of initial spheres as the input of a graph convolution network (GCN) [28, 29, 30, 31] to refine these centers. The pixel features together with the refined centers are then used to compute the new radii. Since there are no connections between the predicted centers up till now, the graph of GCN is represented as an identity matrix.

In the sphere prediction, Chamfer Distance (CD) and Earth Mover's Distance (EMD) are introduced to constrain the sphere centers [14], and Radius (R) loss is used for learning the radii of spheres. CD loss is employed to measure the mismatch between the predicted centers $\mathbf{C_p}$ and the target centers $\mathbf{C_t}$ of the ground truth spheres. EMD loss measures the mismatch between the distributions of sphere centers in the ground truth and the predicted domain. Similar to CD loss, R loss is proposed under the assumption that spheres that are close to each other are more likely to have similar radii, that is,

$$L_r = \sum_{\mathbf{p} \in \mathbf{C_p}} (r_{\mathbf{p}} - r_{\mathbf{p}'})^2 + \sum_{\mathbf{q} \in \mathbf{C_t}} (r_{\mathbf{q}} - r_{\mathbf{q}'})^2, \tag{1}$$

where $r_{\mathbf{p}}$ is the radius of the medial sphere with center $\mathbf{p}$. $\mathbf{p}'$

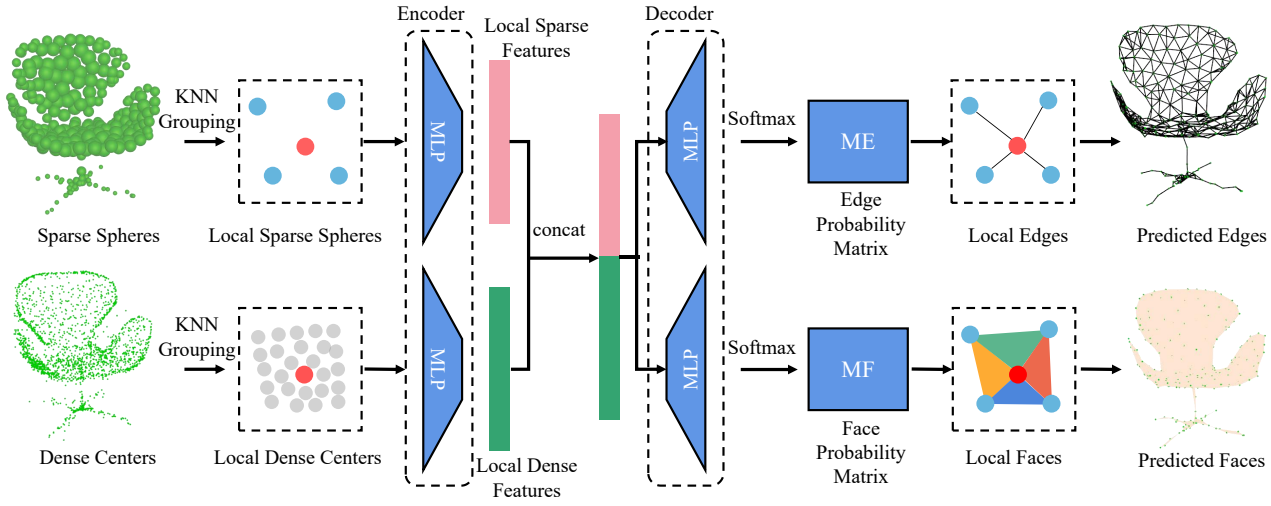

Figure 3: The overall pipeline of the Topology Prediction module.

denotes $\mathbf{p}$'s closest center of target spheres, and its radius is $r_{\mathbf{p}'}$. Accordingly, $\mathbf{q}'$ is the closest predicted center of $\mathbf{q}$.

In the first stage, we consider all of the three losses, *i.e.*,

$$L_{init} = \lambda_1 L_{cd} + \lambda_2 L_{emd} + \lambda_3 L_r. \tag{2}$$

Considering the performance and time consumption of reconstruction, we predict a sparse set of 256 spheres for our MAT representation. Ideally, the spheres should lie on the medial curve and sheets and any outliers of the sparse spheres will cause bumpy and unsmooth structures in the reconstructed mesh after topology prediction. To reduce the outliers, we also predict a dense set of 2048 sphere centers to represent the finely sampled medial curves and sheets, without using the radii.

In general, this module predicts multi-resolution sphere sets. Sparse medial spheres are used to guide the topology prediction, while dense centers will provide richer information for better topology prediction. We will introduce how the dense set is used for refining the topology in the Topology Prediction module.

3.2. Topology Prediction

The topology of a medial mesh is represented by a two-dimensional edge adjacency matrix and a three-dimensional face adjacency matrix. Each element of the adjacency matrices is either 1 or 0 indicating whether there is an edge between two spheres or a face among three spheres. Under this context, edge/face prediction can be regarded as a binary classification task.

However, the full adjacency matrices are sparse, resulting in an unbalanced distribution of 0 and 1, which makes it impossible to achieve a meaningful binary classification.

The topology is related to the euclidean distance of medial spheres, *i.e.*, spheres that are close to each other are more likely to be connected. Therefore, for edge prediction, we only predict the probabilities of medial edges for each sphere and its K nearest neighbors, alleviating the imbalance of the classification labels. For face prediction, we predict the probability of a medial face for each sphere when the other two spheres of the face belong to its K nearest neighbors.

The core idea is to split the global topology into N local topologies, one for each sphere as illustrated as the red sphere in Fig. 3. It is observed that features extracted from sparse spheres are not enough to predict the precise topology. Therefore, the dense centers predicted in the Image2Sphere stage together with the sparse spheres are used to leverage the neighborhoods at multiple scales for achieving both detail capture and prediction robustness. As shown in Fig. 3, two local features extracted from two distinct resolutions (sparse as 256 and dense as 2048 in our experiment) are concatenated to predict the probability of edges and faces. Edges and faces with probability larger than a user-defined threshold ϵ where $0 \leq \epsilon \leq 1$ are selected to construct the topology of the shape.

The loss function of the topology prediction module is the sum of the cross-entropy loss of two binary classifiers,

$$L_{tp} = -\sum_{\mathbf{e}\in\mathbf{ME}} y_{\mathbf{e}} \log(p(\mathbf{e})) - \sum_{\mathbf{f}\in\mathbf{MF}} y_{\mathbf{f}} \log(p(\mathbf{f})), \tag{3}$$

where **ME** and **MF** are probability matrices of edges and faces, respectively. $y_{\mathbf{e}}$ and $y_{\mathbf{f}}$ are the label value (0 or 1) of edge or face and $p(\cdot)$ is the corresponding softmax probability. In this way, three types of medial primitives can be predicted: medial spheres, medial edges, and medial faces.

3.3. MAT Smoothing

After topology prediction, the MAT of the object is obtained and is sufficient to reconstruct its surface mesh. However, such mesh may have an uneven surface second column in Fig. 8) due to the inconsistent distribution of coordinates and radii between connected spheres. Our MAT smoothing module is designed to solve this problem with two stages: smoothing and refinement. Following Eq. (4), for a sphere s, we compute the centroid of all its connected spheres $C(s)$, with $|C(\cdot)|$ being the number of connected spheres. We use a specified smoothing weight $t \in [0, 1]$ to balance the performance of smoothness. The smoothing operation is quite related to [32, 33]. The smaller t is, a stronger smoothing effect is achieved:

$$s_{smooth} = t \cdot s + \frac{(1-t)}{|C(s)|} \sum_{s'\in C(s)} s'. \tag{4}$$

Even though the smoothing operation makes a better visual effect with a smoother surface, it can also shrink the shape by its nature. To maintain the distribution of the coordinates and radii of the medial spheres after smoothing, we train a MAT refinement network that has a similar network structure with the

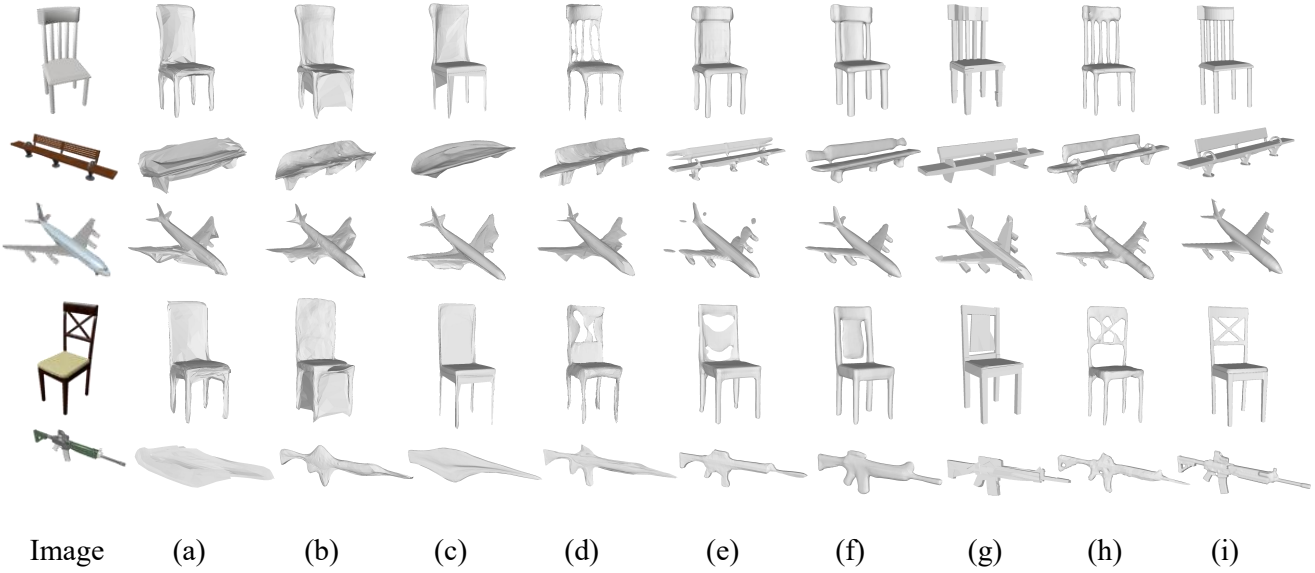

Figure 4: **Qualitative results on mesh reconstruction.** (a) AtlasNet; (b) Pixel2Mesh; (c) TMNet; (d) SkeletonBridge; (e) DISN; (f) OccNet; (g) BSP-Net; (h) Ours; (i) Ground Truth. For the models with holes which are difficult to reconstruct, our method can predict much better results.

previously mentioned sphere refinement network in Sec. 3.1, but with three differences: 1) The input is the smoothed spheres, not their initial centers. 2) The output is the displacements of the coordinates and radii instead of new spheres. 3) The topological relationship is used to support the graph convolution network here, while in the sphere refinement network of Sec. 3.1 there is no connection between spheres.

3.4. Surface Reconstruction From Medial Mesh

The enveloping surface of each medial primitive can be constructed from the union of simpler objects. Since a medial edge is a linear interpolation of two end spheres, its volume is a union of one cone and two spheres. Similarly, the volume of a medial face is a union of three spheres, three cones, and one triangular prism. These characteristics inspire us to use Constructive Solid Geometry (CSG) [34] for surface reconstruction from a medial mesh. We use the VDB data structure [35, 36], a compact volumetric data structure, to achieve high-quality CSG operation of medial primitives. After converting existing medial primitives to implicit level sets, the VDB data structure can perform nearly real-time union operations on these level sets. Finally, the resulting volume is converted to a triangle mesh through the Marching Cubes algorithm [37].

4. Experiments

Dataset

We evaluate our approach on ShapeNet [1] dataset. To the best of our knowledge, computing medial axis transform often needs roughly uniformly distributed, manifold, and closed triangular mesh. But in ShapeNet, lots of meshes are non-manifold or have other problems. Consequently, Q-MAT[27] can not compute medial axis transform of all the shapes in ShapeNet. We finally generated MATs of 47.5% of the full set and named the generated dataset as ShapeNet-MAT. The dataset includes 17,507 samples in 13 categories and the samples are randomly split into two subsets, 80% of samples are used for training and the remaining for testing. Each sample has 24 images with different views provided by [2]. For a fair comparison, all compared methods are re-trained on the same samples.

Implementation details

All networks are trained separately. The Image2Sphere predicts 256 sparse spheres and 2048 dense centers. We use a learning rate of $1e^{-4}$ for the sphere prediction of Image2Sphere. In the first N_1 training epochs, the sub-network using the global feature is trained and then fix their parameters. In the next N_2 epochs, the sub-network using local features is trained.

In the Topology Prediction network, we select 8 neighbors from sparse spheres and 64 neighbors from dense centers for each sphere in the sparse set. We train the Topology Prediction module using a learning rate of $1e^{-3}$. The smoothing weight t is 0.5. The networks are trained individually for each category. We use OpenVDB [36] for implementation of surface reconstruction from MAT. For IoU computation, the resolution of voxel is $32 \times 32 \times 32$. Before triangle mesh generation, we predict topology again and fill the surface holes [13].

4.1. Comparisons with State-of-the-Arts

In this section, qualitative and quantitative comparisons with several state-of-the-art methods for mesh reconstruction, including AtlasNet [15], Pixel2Mesh [14], TMNet [17], SkeletonBridge [10], DISN [20], OccNet [38], BSP-Net [39] are conducted to demonstrate the effectiveness of our MAT-based reconstruction. All methods are trained/tested on the same samples and use their corresponding supervision data representation. In our method, the supervision data is MAT spheres, edges and faces. For AtlasNet, Pixel2Mesh, TMNet, triangular meshes are used for supervision. SkeletonBridge uses three representations: skeleton points, voxel, and triangular mesh. BSP-Net, DISN, and OccNet also use their corresponding supervisory data. In local image feature capture, ground truth camera parameters are used for all methods.

Qualitative results

The qualitative results are shown in Fig. 4. The results show that mesh deformation based methods, *i.e.*, AtlasNet [15], Pixel2Mesh [14], and TMNet [17] can only reconstruct mesh with a similar overall shape but fail to reconstruct topologically complex shapes. Although TMNet eliminates incorrect faces, it

CD ↓	plane	bench	chair	rifle	table	lamp	boat	couch	car	display	phone	speaker	cabinet	mean
AtlasNet	1.32	4.54	5.50	3.65	5.99	14.22	2.66	3.85	1.42	**3.51**	1.21	**10.29**	2.81	4.69
P2M	2.78	6.17	6.40	4.75	4.44	10.28	5.47	4.62	2.02	5.28	1.85	18.56	9.74	6.34
TMNet	2.09	5.56	4.06	2.65	3.68	10.32	4.97	4.21	1.70	6.16	1.05	13.01	**2.62**	4.78
Skeleton	1.44	4.14	3.81	2.33	3.50	8.22	3.19	4.03	2.91	4.16	1.96	12.21	3.76	4.28
DISN	2.08	5.88	5.56	2.81	5.34	13.89	**2.47**	**2.98**	**1.30**	5.25	1.29	16.26	4.18	5.33
OccNet	1.48	5.35	4.36	2.47	5.14	9.51	4.23	5.22	1.91	6.85	0.96	16.58	5.71	5.37
BSP-Net	1.41	4.79	4.59	2.91	4.75	10.74	4.72	4.85	1.69	6.48	1.36	14.93	4.33	5.20
Ours	**1.24**	**3.66**	**3.25**	**1.61**	**3.23**	**6.47**	2.95	3.45	1.36	4.25	**0.93**	11.56	4.12	**3.70**
IoU ↑	plane	bench	chair	rifle	table	lamp	boat	couch	car	display	phone	speaker	cabinet	mean
AtlasNet	0.546	0.429	0.388	0.443	0.458	0.326	0.440	0.405	0.558	0.445	0.652	**0.290**	**0.467**	0.450
P2M	0.302	0.417	0.398	0.524	0.482	0.354	0.435	0.408	0.545	0.416	0.610	0.252	0.279	0.417
TMNet	0.493	0.360	0.405	0.492	0.495	0.358	0.411	0.437	0.555	0.439	0.667	0.270	0.455	0.449
Skeleton	0.504	0.448	0.400	0.533	0.489	0.377	0.450	0.411	0.445	0.443	0.605	0.279	0.394	0.444
DISN	0.501	0.435	0.379	0.524	0.533	0.305	**0.461**	**0.444**	**0.581**	0.423	0.655	0.259	0.391	0.453
OccNet	**0.591**	0.477	0.434	0.521	**0.541**	0.303	0.454	0.431	0.524	0.417	0.671	0.254	0.379	0.461
BSP-Net	0.555	0.469	0.394	0.485	0.475	0.315	0.393	0.392	0.550	0.412	0.597	0.246	0.389	0.436
Ours	0.558	**0.511**	**0.446**	**0.612**	0.510	**0.387**	0.434	0.415	0.540	**0.448**	**0.675**	0.257	0.388	**0.475**
EMD ↓	plane	bench	chair	rifle	table	lamp	boat	couch	car	display	phone	speaker	cabinet	mean
AtlasNet	2.38	3.41	4.18	3.37	3.93	5.69	2.89	3.00	2.14	2.88	2.01	4.20	2.91	3.31
P2M	2.99	4.11	4.74	4.19	3.35	5.91	4.41	3.26	2.63	3.66	2.20	6.68	5.39	4.12
TMNet	2.75	3.50	4.03	3.39	3.15	5.98	3.50	2.62	2.18	2.99	**1.49**	**4.14**	**2.36**	3.24
Skeleton	2.52	4.12	3.79	3.95	3.32	6.23	3.67	3.15	3.55	3.21	2.59	5.42	3.73	3.79
DISN	2.75	3.13	3.61	3.21	2.98	6.46	**2.56**	**2.43**	**2.11**	3.02	1.71	4.41	2.95	3.18
OccNet	**2.03**	2.97	3.18	3.01	2.90	5.01	3.27	2.91	2.30	3.09	1.52	4.70	3.24	3.09
BSP-Net	2.21	3.27	3.91	3.53	3.04	6.30	3.81	3.16	2.83	3.11	1.96	4.88	3.08	3.46
Ours	2.13	**2.80**	**3.05**	**2.53**	**2.79**	**4.93**	2.88	2.55	2.13	**2.76**	1.56	4.28	2.96	**2.87**
F-score ↑	plane	bench	chair	rifle	table	lamp	boat	couch	car	display	phone	speaker	cabinet	mean
AtlasNet	94.07	77.61	72.03	82.29	78.78	60.15	84.11	78.63	92.29	**83.86**	93.96	**60.13**	78.52	79.73
P2M	86.85	73.55	64.02	80.12	83.24	66.05	73.61	72.55	86.95	74.31	88.02	48.83	53.44	73.20
TMNet	91.64	72.81	78.90	86.05	87.23	67.71	71.58	80.35	90.99	77.48	95.05	55.56	**82.29**	79.82
Skeleton	93.29	79.57	78.62	89.22	83.73	68.79	80.43	77.89	79.84	79.82	87.43	57.07	73.06	79.14
DISN	92.26	79.44	73.46	87.65	82.39	68.81	**86.44**	**82.43**	**93.13**	79.98	93.29	53.74	71.88	80.38
OccNet	93.69	83.09	80.47	88.45	82.91	68.54	80.29	75.96	88.82	75.43	95.78	53.28	66.53	79.48
BSP-Net	91.78	80.29	77.10	84.35	84.22	68.09	80.22	76.33	91.10	76.99	93.59	51.60	73.70	79.18
Ours	**94.96**	**84.42**	**83.70**	**92.40**	**87.33**	**74.32**	83.20	79.97	92.41	81.81	**96.12**	56.18	72.95	**83.06**

Table 1: **Quantitative results on mesh reconstruction.** The Chamfer Distance, IoU, Earth Mover's Distance, and F-score are used. The best results are **boldfaced**, and the second best results are underlined.

is still constrained by the topology of the initial spherical mesh, as illustrated by the chair and bench cases in the first two rows. SkeletonBridge [10] directly predicts skeleton points of the 3D shape, which makes it possible to generate topologically complex shapes. However, the fine details of the object still cannot be fully captured, such as the chair back and airplane propellers. Implicit methods [20, 38] are capable to reconstruct smooth meshes, but there is a gap in thickness between the shape and the ground truth, as shown in Fig. 7. BSP-Net [39] directly extracts a polygon mesh via convex decomposition and recovers sharp geometric details. However, there are many overlapping faces inside its mesh, and the surface lacks smoothness. Our method has achieved better performances as demonstrated by the similar overall shape as well as the exquisite local details. Besides, with the conversion of MAT to VDB implicit surface representation, our generated surface meshes are guaranteed to be manifold and watertight, without any self-intersection.

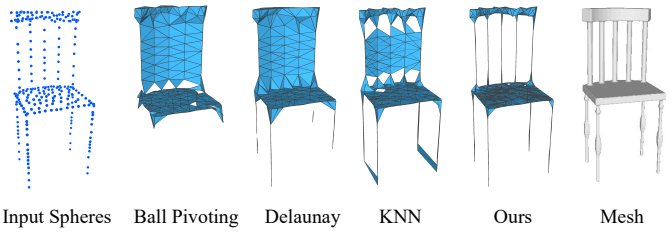

Figure 5: **Qualitative results of topology generation** methods by connecting vertices to form edges and faces.

Quantitative results

We adopt the widely used Chamfer Distance (CD) loss, Earth Mover's Distance (EMD) loss, F-score [14], Intersection over Union (IoU) of the voxels as comparison metrics. After aligning the prediction results with ground truth, 10,000 points are uniformly sampled from each triangle mesh. The metrics are calculated on the sampled points and the vertices of ground truth meshes. Since the training and testing samples are less than the full dataset, all methods have a certain decline in quantitative than using the full dataset, but because all methods use the same samples, the performance difference between methods remains. For example, OccNet and DISN are better than Pixel2Mesh quantitatively. Table 1 shows our approach outperforms the state-of-the-art methods in all metrics over most categories. It is noticed that all of the mesh-based methods take meshes as supervision information in the training process, which aims to directly minimize the losses calculated on meshes. Our method learns the medial spheres and topological relations without using the surface meshes as supervision information, but still achieves better (or comparable) results on the reconstruction error of reconstructed meshes.

Topology prediction

Given sparse medial spheres, we compare the topology generation with alternative methods: Ball pivoting [40], Delaunay tri-

Figure 6: Test results on real images.

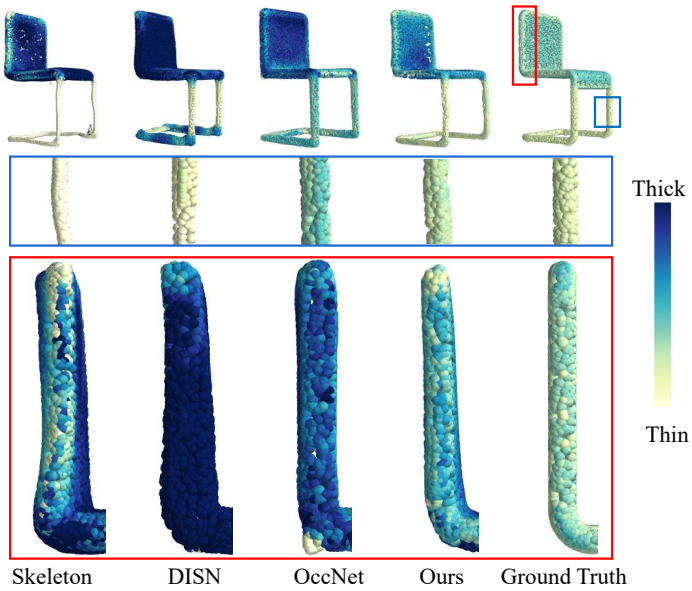

Figure 7: **Comparison on Shape Diameter Function.** The color of the point indicates its thickness.

angulation (deleting overlong edges using a threshold), and K Nearest Neighbor (KNN) (connecting the K nearest neighbors of each sphere and extracting the formed faces). Fig. 5 shows that the alternative methods cannot generate the complex topology correctly (especially linear structures). Our Topology Prediction module predicts the delicate back and legs of the chair.

Testing on real image

We test our model on real images from the Pix3d [41] dataset. Although our model is only trained on the ShapeNet dataset, it generalizes well to real-world objects (Fig. 6). Our method has the ability to reconstruct the hole structure of 3D shapes.

Comparison of Shape Diameter Function (SDF)

We apply Shape Diameter Function (SDF) [42] to measure the local thickness of reconstructed meshes.

We first compute the SDF value for each face of mesh and sample M (10K in our experiments) points in total on these faces. The SDF value η of each point corresponds to the face from which it is sampled. A larger SDF value indicates thicker volume below the surface point.

We propose the average thickness error to measure the difference of the local thickness between the predicted mesh and the ground truth. To the best of our knowledge, none of the traditional quantitative metrics could reflect the thickness error of the local shape. For a point $\mathbf{p}$ on the ground truth mesh surface, we find the nearest point $\mathbf{p}'$ from the predicted mesh surface as its corresponding point.

The absolute value of difference between the thickness values of $\mathbf{p}$ and $\mathbf{p}'$ is $|\eta_{\mathbf{p}} - \eta_{\mathbf{p}'}|$. Similar to the R loss, we calculate the thickness error in both directions, *i.e.*,

$$\eta_e = \frac{1}{2M}\left(\sum_{\mathbf{p}} |\eta_{\mathbf{p}} - \eta_{\mathbf{p}'}| + \sum_{\mathbf{q}} |\eta_{\mathbf{q}} - \eta_{\mathbf{q}'}|\right), \tag{5}$$

where $\mathbf{q}$ and $\mathbf{q}'$ are the point on the ground truth mesh and its nearest point on the predicted mesh, respectively.

The computation of SDF needs closed and manifold mesh with correct normal information, but it could not be guaranteed that the predicted meshes of the mesh-based methods we compare have these attributes. We compare 4,573 samples of which the SDFs are successfully computed. The qualitative and quantitative results show our results have a closer thickness to the ground truth than other methods. Table 2 shows the average thickness errors (lower is better) of the reconstructed meshes and the ground truth.

Our result is closer to the ground truth on thickness than the compared methods. The visual comparison of SDF in Fig. 7 also gives the same conclusion.

Category	Skeleton	DISN	OccNet	Ours
plane	1.811	1.867	1.796	**1.607**
chair	1.874	1.674	1.619	**1.492**
firearm	1.638	1.008	1.095	**0.985**
table	1.618	1.235	1.285	**1.226**
mean	1.774	1.516	1.503	**1.361**

Table 2: Quantitative comparison on average thickness error.

High genus shape comparison

To show the effectiveness of our method on topologically complex shapes, we compare the results of higher-genus samples on 5 categories (bench, chair, firearm, plane, table), containing a total of 29,520 samples, with average genera of 11.4 per sample. As shown in Table 3, our results still perform well.

Method	CD	IoU
AtlasNet	4.753	0.435
P2M	5.519	0.451
TMNet	3.904	0.436
SkeletonBridge	3.317	0.465
DISN	4.175	0.489
OccNet	3.936	0.496
BSP-Net	4.064	0.459
Ours	**2.902**	**0.515**

Table 3: Quantitative comparisons on high-genus samples.

4.2. Ablation Study and Application

Effect of each stage on sphere prediction

The initial spheres decoded from the global image features only achieve the similarity of the overall shape, so we improve it with sphere refinement and MAT Smoothing. Table 4 and Fig. 8 show the effect of each stage quantitatively and visually.

The sphere refinement in Image2Sphere module generates complex local shape details and the MAT Smoothing module produces clearer boundaries of the shape and more consistent radii of neighboring spheres. In the smooth operation, we use the weight $t = 0.5$ for Eq. (4). We calculate the ratio of diagonal of the bounding box between current spheres and real data,

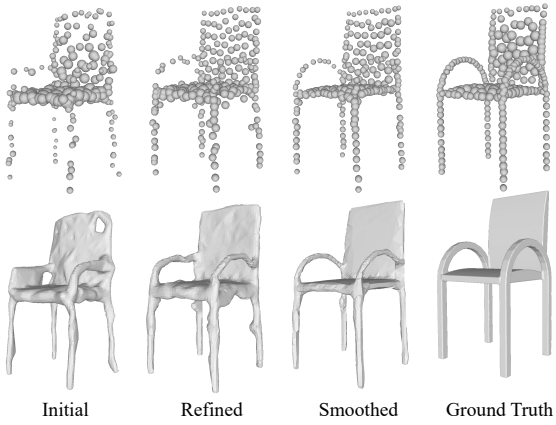

Figure 8: Visual comparisons on spheres and reconstructed meshes of all stages. The two rows are the predicted spheres and the corresponding mesh reconstructed after topology prediction. Although the spheres learned from global and local features can learn the geometric characteristics of the object, there is still a certain gap with the ground truth. After smoothing and learning by the topological information, the spheres have been significantly optimized on the curve and plane.

Stage	CD (Sphere)	R (Sphere)	CD (Mesh)
Initial	129.28	226.75	5.06
Refined	124.63	196.87	4.38
Smoothed	**124.38**	**191.64**	**3.70**

Table 4: Quantitative comparisons of the three stages.

and the average value in 13 categories is 0.946. Through the following refinement, this ratio is restored to 0.984, indicating that MAT Smoothing not only smooths the surface but also keeps the size of the shape.

Thresholds of topology prediction

In the topology prediction, the probability threshold determines the edges/faces prediction and thus affects the mesh reconstruction results. In Fig. 9, we visualize the reconstruction results of different thresholds and explore the selection strategy. Low edge/face thresholds (τ_e/τ_f) result in local redundancy of the mesh, while high thresholds may lead to incomplete local shape. We select the balanced thresholds ($\tau_e = 0.2, \tau_f = 0.3$) for topology prediction.

Dense centers effect in Topology Prediction

As shown in Fig. 10, although using 256 spheres for topology prediction alone can predict most of the edges and faces correctly, it may lead to incorrect local predictions, such as redundancy or loss of local connections.

Application on topology-guided segmentation

MAT's topology is the abstraction of the 3D shape. Although our task is shape reconstruction, we find that with the guidance of the predicted topology of MAT, the overall shape can be easily segmented into multiple parts without labeling and deep learning. Fig. 12 visualizes the segmentation results on meshes. It can be seen that the surface structure and curve structure can be clearly distinguished on the reconstructed mesh. With the help of the predicted topology, we can decompose the shape into multiple parts without ground truth labels or prediction. As shown in Fig. 11, the topology-guided segmentation consists of three steps:

1) Semantic Split. Intuitively, a sphere on a surface has more faces than a sphere on a curve. According to the number of faces,

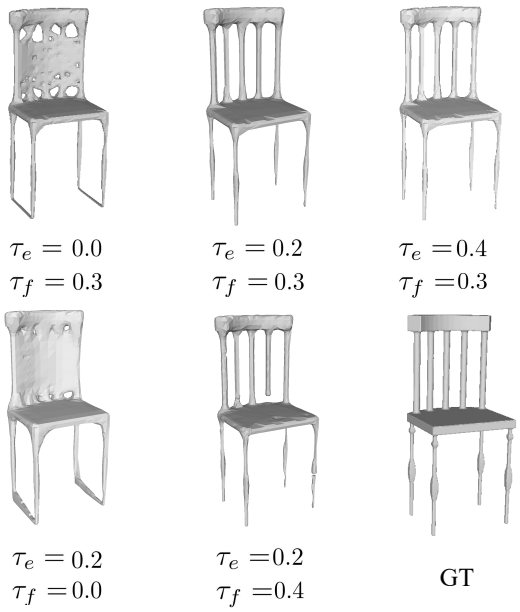

Figure 9: Influence of edge/face probability threshold τ_e/τ_f on mesh reconstruction.

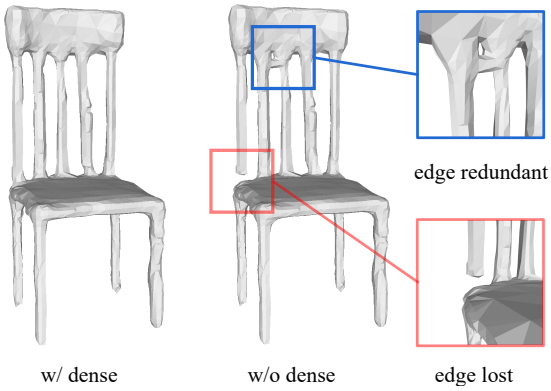

Figure 10: Reconstruction results with and without dense centers in topology prediction.

we can split the spheres into two semantics: curve or surface. The sphere with the number of prediction faces greater than the threshold K is be seen as on the surface, otherwise on the curve.

2) Parts Clustering. We use DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to cluster curve spheres and surface spheres. For the curve spheres, we cluster them based on 3D spatial coordinates. For surface spheres, we compute the average normal of the faces those the sphere is located in as its normal (absolute value) and cluster them by using the coordinates and normals, to distinguish the connected surfaces with different normals, such as the back and base of a chair.

3) Mesh Correspondence. After splitting the parts of MAT spheres, we segment the triangle faces by finding their nearest spheres. Finally, we split the triangular mesh into many parts.

Although our method generates overall shape from a single view image, the triangular mesh can easily be segmented into parts based on MAT segmentation. Due to the DBSCAN needs not specifying the number of cluster centers, the parts number of the shape is adaptive.

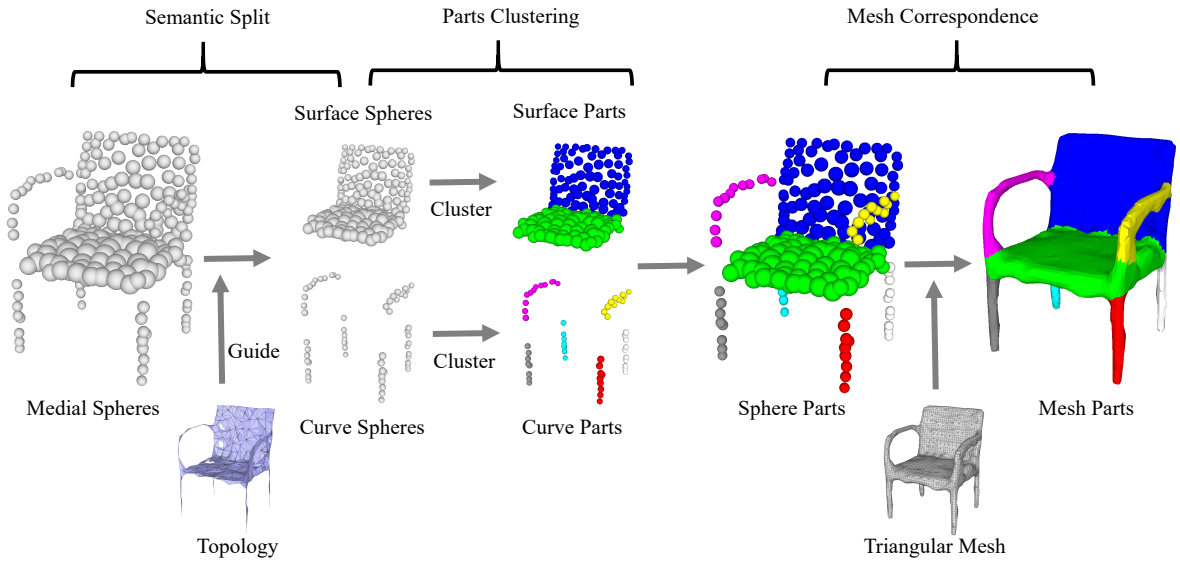

Figure 11: Topology-guided segmentation on predicted MAT and predicted triangular mesh.

More samples in all stages

o show the effect of each stage, the output results of all stages are shown in Fig. 13 including sphere prediction, topology prediction, MAT smoothing, and mesh reconstruction.

We also show the thickness information (shape diameter function) of reconstruction results in different categories, which is not considered by other methods. It can be seen that our method can gradually generate fine spheres on a skeleton, and reasonably predict the topological relationship of the spheres. Finally, construct a complete MAT representation and reconstruct a complex surface mesh. We also show the local thickness that is affected by the radius through shape diameter function.

Network architectures

Fig. 14 shows the architecture of initial sphere prediction. The input image is encoded to a global vector by ResNet18. Then the global feature vector is input to three sub-networks. Two networks encode the feature and decode it into initial centers and radii, and then the feature vector is input to the displacement learning module to learn the displacements for the initial centers. The displacement learning module follows 3DN [16]. The architecture of sphere refinement network (Fig. 16) follows Pixel2Mesh. We make the following changes in implementation: 1) The input of deformation is 256 center coordinates. 2) The input graph is an identity matrix for graph convolution and graph decoder. 3) The output of the last layer is the refined radii of 256×1. 4) The R loss is added to the loss function. In Image2Sphere, the initial sphere prediction uses a batch size of 24, and the refinement network uses a batch size of 1. The learning rate is $1e^{-4}$. In the topology prediction network (Fig. 15), we group 8 neighbor spheres from sparse spheres and 64 neighbor centers from dense centers for each sparse sphere. We only predict the topology of sparse spheres.

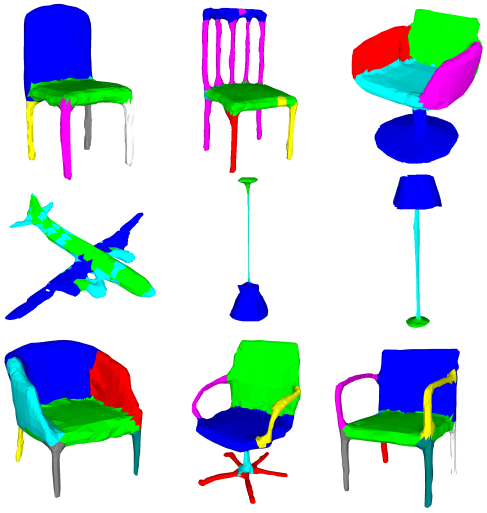

Figure 12: MAT topology-guided parts segmentation results.

5. Conclusion and Future Work

We introduce IMMAT, the first supervised method to learn MAT from a single view image to reconstruct surface mesh. The predicted MAT contains both geometry (spheres) and topology (edges and faces) information, which helps us generate a complex surface mesh that is manifold and watertight. Different from the skeleton points, MAT is a "complete" shape descriptor that can be directly used to reconstruct the surface.

Compared with the state-of-the-art methods, meshes generated by IMMAT exhibit superior visual quality and have more accurate local thickness information. All the results show that predicting the MAT inside the shape to recover surface mesh is worthy of further exploration. It has no topological constraints and can generate complex shapes. With the help of predicted MAT, we can easily segment the parts of the reconstructed mesh without any supervision, which other representations may not accomplish. There are two research directions worth exploring in the future: exploring the applications of MAT in non-rigid shape reconstruction from single view images, and self-supervised learning of MAT by differentiable rendering.

Acknowledgements

This work was supported by the National Key R&D Program of China (2020YFB1708900) and the China National Natural Science Foundation (62072271).

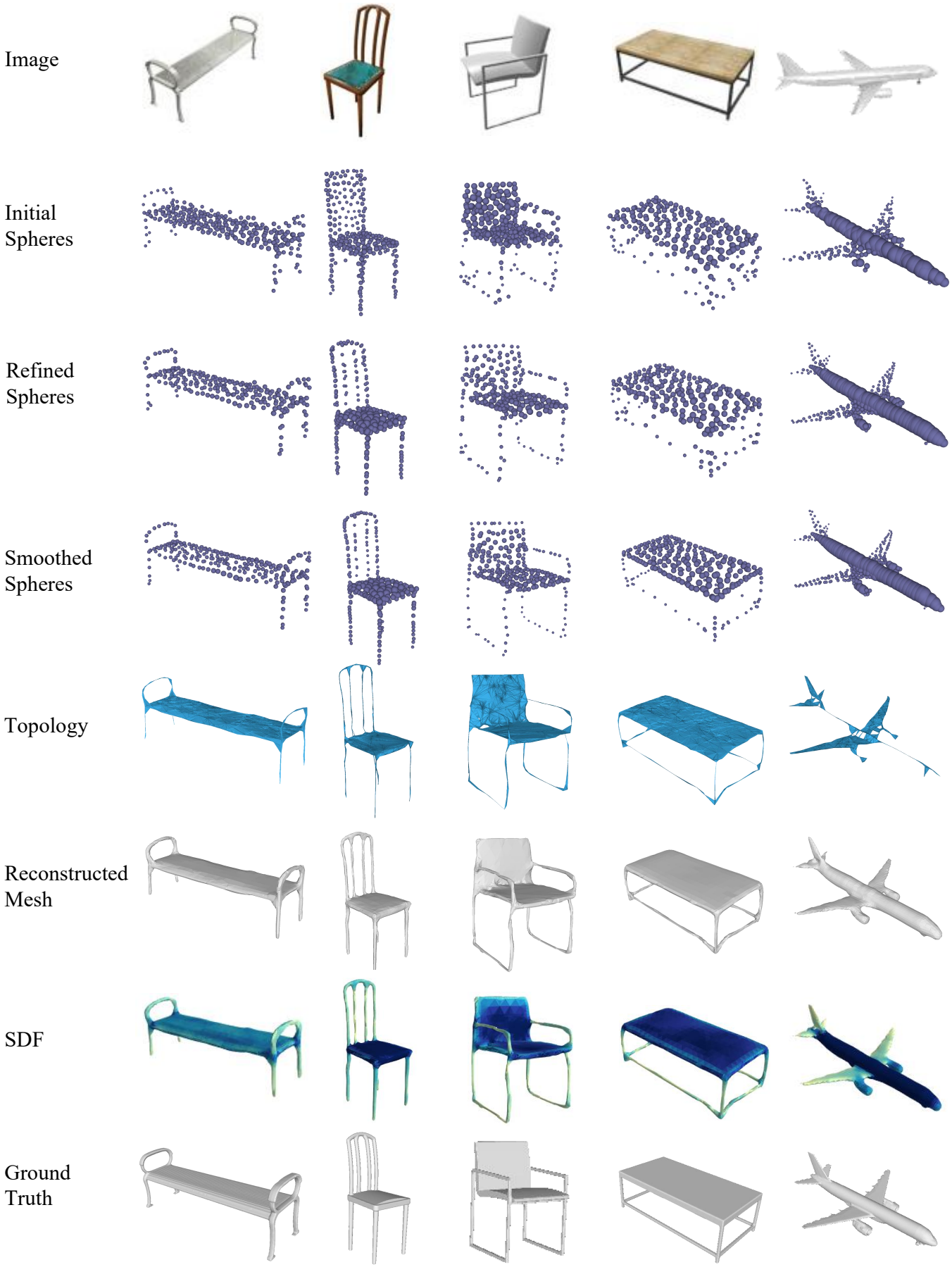

Figure 13: Visualization results of all stage outputs.

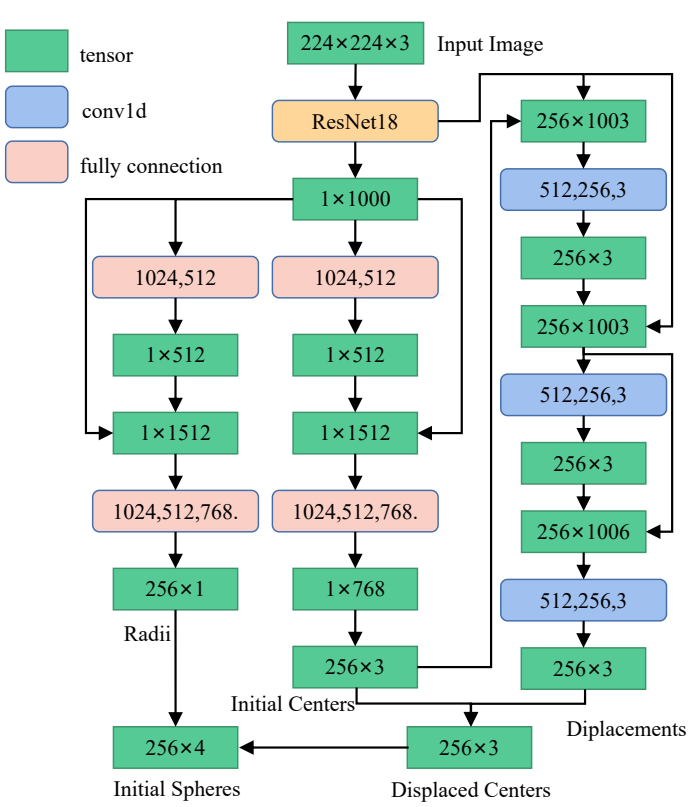

Figure 14: Initial sphere prediction architecture.

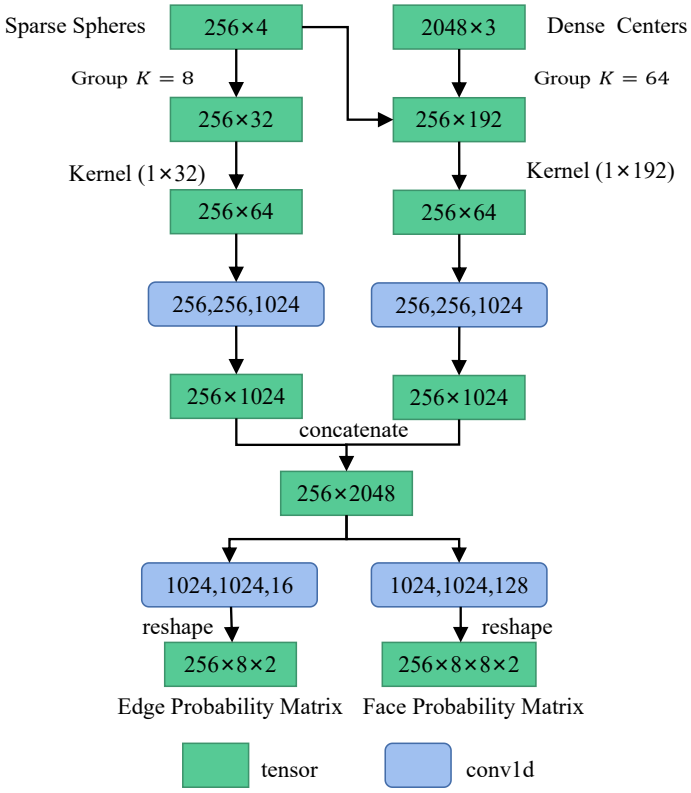

Figure 15: Topology prediction architecture.

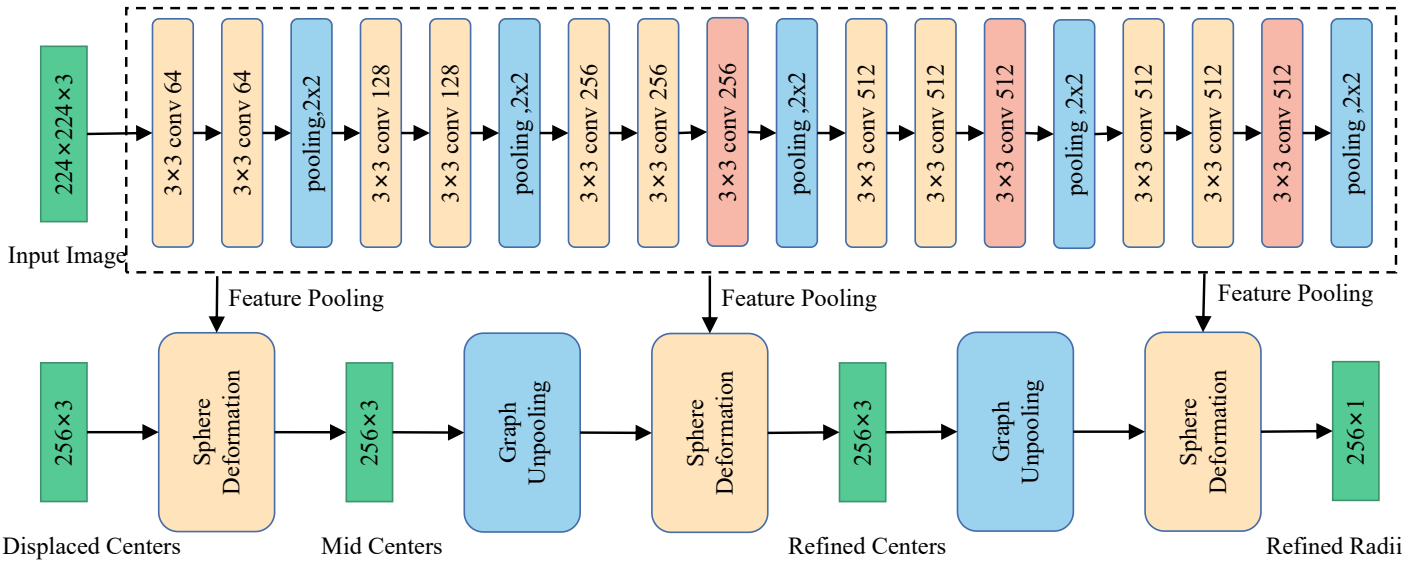

Figure 16: Sphere Refinement Network.

References

[1] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, Shapenet: An information-rich 3d model repository, Computer Science.

[2] C. B. Choy, D. Xu, J. Y. Gwak, K. Chen, S. Savarese, 3D-R2N2: A unified approach for single and multi-view 3d object reconstruction, in: European Conference on Computer Vision, 2016.

[3] S. Tulsiani, T. Zhou, A. A. Efros, J. Malik, Multi-view supervision for single-view reconstruction via differentiable ray consistency, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2626–2634.

[4] S. Yamaguchi, S. Saito, K. Nagano, Y. Zhao, W. Chen, K. Olszewski, S. Morishima, H. Li, High-fidelity facial reflectance and geometry inference from an unconstrained image, ACM Transactions on Graphics (TOG) 37 (4) (2018) 1–14.

[5] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, C. Schmid, Bodynet: Volumetric inference of 3d human body shapes, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 20–36.

[6] X. Han, Z. Li, H. Huang, E. Kalogerakis, Y. 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, 2017, pp. 85–93.

[7] Y. Yang, C. Feng, Y. Shen, D. Tian, Foldingnet: Point cloud auto-encoder via deep grid deformation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 206–215.

[8] H. Fan, H. Su, L. J. Guibas, A point set generation network for 3d object reconstruction from a single image, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 605–613.

[9] M. Michalkiewicz, J. K. Pontes, D. Jack, M. Baktashmotlagh, A. Eriksson, Implicit surface representations as layers in neural networks, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[10] J. Tang, X. Han, J. Pan, K. Jia, X. Tong, A skeleton-bridged deep learning approach for generating meshes of complex topologies from single rgb images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4541–4550.

[11] H. Blum, et al., A transformation for extracting new descriptors of shape, Vol. 4, MIT press Cambridge, 1967.

[12] W. E. Lorensen, H. E. Cline, Marching cubes: A high resolution 3d surface construction algorithm, ACM siggraph computer graphics 21 (4) (1987) 163–169.

[13] C. Lin, C. Li, Y. Liu, N. Chen, Y.-K. Choi, W. Wang, Point2skeleton: Learning skeletal representations from point clouds (2021) 4277–4286.

[14] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, Y.-G. Jiang, Pixel2mesh: Generating 3d mesh models from single rgb images, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[15] T. Groueix, M. Fisher, V. G. Kim, B. Russell, M. Aubry, AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation, in: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2018.

[16] W. Wang, D. Ceylan, R. Mech, U. Neumann, 3DN: 3d deformation network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[17] J. Pan, X. Han, W. Chen, J. Tang, K. Jia, Deep mesh reconstruction from single rgb images via topology modification networks, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2020.

[18] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, A. Geiger, Occupancy networks: Learning 3d reconstruction in function space, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[19] J. J. Park, P. Florence, J. Straub, R. Newcombe, S. Lovegrove, Deepsdf: Learning continuous signed distance functions for shape representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[20] Q. Xu, W. Wang, D. Ceylan, R. Mech, U. Neumann, Disn: Deep implicit surface network for high-quality single-view 3d reconstruction, in: Advances in Neural Information Processing Systems, 2019, pp. 492–502.

[21] K. Genova, F. Cole, D. Vlasic, A. Sarna, T. Funkhouser, Learning shape templates with structured implicit functions.

[22] K. Genova, F. Cole, A. Sud, A. Sarna, T. A. Funkhouser, Deep structured implicit functions.

[23] K. Genova, F. Cole, A. Sud, A. Sarna, T. Funkhouser, Local deep implicit functions for 3d shape, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4857–4866.

[24] A. Tagliasacchi, T. Delame, M. Spagnuolo, N. Amenta, A. Telea, 3d skeletons: A state-of-the-art report, in: Computer Graphics Forum, Vol. 35, Wiley Online Library, 2016, pp. 573–597.

[25] J. Hu, B. Wang, L. Qian, Y. Pan, X. Guo, L. Liu, W. Wang, Mat-net: Medial axis transform network for 3d object recognition., in: IJCAI, 2019, pp. 774–781.

[26] B. Yang, J. Yao, B. Wang, J. Hu, Y. Pan, T. Pan, W. Wang, X. Guo, P2mat-net: Learning medial axis transform from sparse point clouds, Computer Aided Geometric Design 80 (2020) 101874.

[27] P. Li, B. Wang, F. Sun, X. Guo, C. Zhang, W. Wang, Q-mat:computing medial axis transform by quadratic error minimization, ACM Transactions on Graphics 35 (1) (2015) 1–16.

[28] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, Advances in neural information processing systems 29 (2016) 3844–3852.

[29] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907.

[30] D. Boscaini, J. Masci, E. Rodolà, M. Bronstein, Learning shape correspondence with anisotropic convolutional neural networks, Advances in neural information processing systems 29.

[31] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE transactions on neural networks 20 (1) (2008) 61–80.

[32] A. Tagliasacchi, M. Olson, H. Zhang, G. Hamarneh, D. Cohen-Or, Vase: Volume-aware surface evolution for surface reconstruction from incomplete point clouds, in: Computer Graphics Forum, Vol. 30, Wiley Online Library, 2011, pp. 1563–1571.

[33] A. Tagliasacchi, I. Alhashim, M. Olson, H. Zhang, Mean curvature skeletons, in: Computer Graphics Forum, Vol. 31, Wiley Online Library, 2012, pp. 1735–1744.

[34] M. Douze, J.-S. Franco, B. Raffin, Quickcsg: Arbitrary and faster boolean combinations of n solids.

[35] K. Museth, Vdb: High-resolution sparse volumes with dynamic topology, ACM transactions on graphics (TOG) 32 (3) (2013) 1–22.

[36] K. Museth, J. Lait, J. Johanson, J. Budsberg, R. Henderson, M. Alden, P. Cucka, D. Hill, A. Pearce, Openvdb: an open-source data structure and toolkit for high-resolution volumes, in: Acm siggraph 2013 courses, 2013, pp. 1–1.

[37] G. M. Nielson, Dual marching cubes, in: IEEE Visualization 2004, IEEE, 2004, pp. 489–496.

[38] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, A. Geiger, Occupancy networks: Learning 3d reconstruction in function space, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[39] Z. Chen, A. Tagliasacchi, H. Zhang, Bsp-net: Generating compact meshes via binary space partitioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 45–54.

[40] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, G. Taubin, The ball-pivoting algorithm for surface reconstruction, IEEE transactions on visualization and computer graphics 5 (4) (1999) 349–359.

[41] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, W. T. Freeman, Pix3D: Dataset and methods for single-image 3d shape modeling, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[42] L. Shapira, A. Shamir, D. Cohen-Or, Consistent mesh partitioning and skeletonisation using the shape diameter function, The Visual Computer 24 (4) (2008) 249.