S3DS: Self-supervised Learning of 3D Skeletons from Single View Images

Jianwei Hu
hjw52it@gmail.com
School of Software, Tsinghua University & BNRist
China

Ningna Wang
ningna.wang@utdallas.edu
Department of Computer Science, UT Dallas
USA

Baorong Yang
yangbaorong@jmu.edu.cn
College of Computer Engineering, Jimei University
China

Gang Chen
g-chen21@mails.tsinghua.edu.cn
School of Software, Tsinghua University & BNRist
China

Xiaohu Guo
xguo@utdallas.edu
Department of Computer Science, UT Dallas
USA

Bin Wang*
wangbins@tsinghua.edu.cn
School of Software, Tsinghua University & BNRist
China

ABSTRACT

3D skeleton is an inherent structure of objects and is often used for shape analysis. However, most supervised deep learning methods, which directly obtain 3D skeletons from 2D images, are constrained by skeleton data preparation. In this paper, we introduce a self-supervised method S3DS: a differentiable rendering-based method to reconstruct a 3D skeleton of shape from its single-view images, by using medial axis transformation (MAT) as its 3D skeleton. We use medial spheres (center positions and radii) to represent the 3D skeleton and use the connectivity of the spheres (medial mesh) to represent the topology. We trained a medial sphere prediction network, which reconstructs 3D skeleton spheres (centers and radii) from a single-view image and renders them into a 2D silhouette with many circles. Because of the radius, the center of the circle will fall on the 2D skeleton. Then the 3D spheres are fitted to the 3D skeleton by fitting many 2D circles onto the 2D skeleton. A mechanism is proposed to generate the connectivity of the discrete medial spheres and construct the 3D topology of the shape. We have conducted extensive experiments on public datasets and proved that S3DS has better performance than baseline and competitive performances with supervised methods on 3D skeletons reconstruction.

CCS CONCEPTS

• **Computing methodologies** → **Shape representations**; **Computer vision**; **Shape modeling**.

KEYWORDS

3d skeletons; reconstruction; self-supervised; machine learning

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

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612204

ACM Reference Format:
Jianwei Hu, Ningna Wang, Baorong Yang, Gang Chen, Xiaohu Guo, and Bin Wang. 2023. S3DS: Self-supervised Learning of 3D Skeletons from Single View Images. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3581783.3612204

1 INTRODUCTION

With the extension of deep learning in 3D visual tasks, the reconstruction of 3D representation from 2D images has received massive attention and has been widely used in computer graphics, computer-aided design, automatic driving, virtual reality, and other fields.

In the 3D world, the 3D skeleton of a shape often contains information about the geometric distribution and topology of the shape, which is of great help in shape analysis and reconstruction of shapes with complex topology. Many works are devoted to applying 3D skeleton representation to various 3D vision tasks. MAT-Net [15] applied medial axis transform to the classification task. P2MAT-Net [43] and Point2Skeleton [21] generate the medial skeletons from 3D point clouds in supervised and self-supervised ways respectively. 3D skeleton points could also be used as a medium representation to generate point clouds, voxels, and triangular meshes from 2D images [14, 29, 35, 36].

SkeletonBridge [35] and IMMAT [14] are supervised methods to extract 3D skeletons from images. SkeletonBridge captures the underlying topological structure of the target object and takes it as a bridge in a single-view image reconstruction task, using the corresponding meso-skeleton generated with DPC [40] as supervision.

However, the meso-skeleton is only used as the initial geometry, so the reconstructed results are visually unsatisfactory. IMMAT uses the medial axis transformation (MAT) [4] as the target representation for skeleton learning and achieves state-of-the-art performance. At the same time, IMMAT faces the difficulty of preparing massive MAT data, as the computation of MATs highly depends on the quality of the input CAD models. Therefore, IMMAT only uses 47.5% of samples from the 13 categories of ShapeNet [5] for experiments.

To address the above issues, we propose to predict MAT for precise geometric reconstruction from images with differentiable

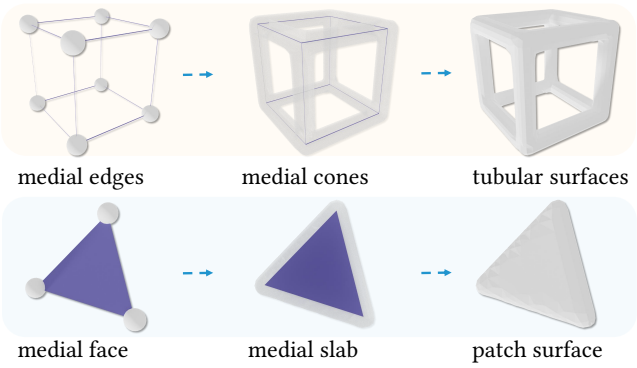

Figure 1: Atomic elements of medial axis transform (medial mesh): spheres, edges, faces. Medial spheres represent the skeleton distribution of a 3D shape.

rendering in a self-supervised manner to overcome the limitation of MAT data preparation.

Figure 1 shows the three elements of MAT. The medial spheres, whose centers are located on the skeleton, represent the maximum inscribed spheres for the given surfaces, and their radii represent the local thickness of the shapes. Compared with the representation of pure skeleton curves [35], our choice of MAT is more suitable for differential rendering (thus more suitable for self-supervised learning) in the following three aspects.

Firstly, skeleton curves have no radius information, and cannot be used to render the 2D silhouette of shapes. On the contrary, the projection of a set of medial spheres, with radii, can provide a holistic 2D silhouette of the shape. Secondly, the 3D medial sphere and its rendered 2D circle are symmetrical, which means the center of the ground-truth 2D circle falls on the 2D medial axis of rendered shape. By restricting the rendered 2D circles within the silhouette of the target image, their centers will be optimized as close to canonical 2D skeletons as possible. Owing to the symmetric property of MAT, as a result, the corresponding 3D medial spheres will be placed at the optimal position while covering the neighboring areas. Thirdly, the final mesh surface can be restored using the connectivity of medial spheres, so called medial mesh. By constructing the edges and faces among the obtained medial spheres, we can form cone and slab structures and reconstruct the mesh surface, as shown in Figure 1.

In this paper, we propose a self-supervised learning method **S3DS**, based on the above MAT advantages, to learn the 3D skeletons of a shape from a single-view image. As shown in Figure 2, we first generate medial spheres through differentiable rendering, then we produce the connectivity (medial mesh) of the generated spheres. The medial mesh also represents the connectivity of the 3D skeleton of the shape.

One long-standing drawback of existing differentiable rendering methods, which use 2D IoU between the rendered image and the input image as the loss function, is that they tend to focus more on reconstructing large patches of the shape, e.g., the backpack and the base of a chair. However, those fine structures, e.g., chair legs and armrests, may be sacrificed in the loss function and omitted in the reconstructed shape. To tackle this difficulty, we propose a *semantic sphere learning module* to split the silhouette of an image into two sub-silhouettes: a fine silhouette and a coarse silhouette. As shown in Figure 2, the fine structure corresponds to the silhouette of lines and frames of a shape, and the coarse structure corresponds to the silhouette of large patches. We learn the medial spheres for these two structures and optimize them according to their 2D silhouettes respectively. Then we render the merged medial spheres to generate a complete silhouette that approximates the whole shape. Using this fine-coarse strategy, our method outperforms existing approaches significantly on models with fine details.

We conduct extensive ablation studies to show the effectiveness of our proposed approach. The contributions of this paper can be summarized as follows:

- We introduce MAT as the representation of 3D skeletons for shape reconstruction from a single-view image, and propose a novel deep learning framework for self-supervised MAT prediction, without 3D skeleton data preparation.
- We propose the *semantic sphere learning module*, the first differentiable method for predicting 3D medial spheres from a 2D image. The proposed module not only learns large patches of the shape but also produces superior results over fine regions.
- We propose a heuristic approach to generate the connectivity of medial spheres to form medial meshes, whose envelope shape reconstructs the 3D surface.

2 RELATED WORKS

2.1 Skeleton and Medial Axis Transform

Q-MAT [19] put forward the concept of simplification of MAT [4] and media mesh, which is composed of medial spheres, edges, and faces, and expresses the internal skeleton of 3D objects. The centers and radii of the spheres represent the location distribution and volume distribution respectively. Many recent works [37, 42] attempt to generate better quality MAT for various applications. IMMAT [14] reconstructed MAT from a single view image, then constructed the surface triangular mesh of the object, achieving great reconstruction performance in complex shapes. However, it needs to compute MATs from densely-sampled manifold meshes using existing MAT generation methods as 3D supervision, which takes much effort in the data preparation. MAT-Net[15] and P2MAT-Net[43] only use 83.2% MAT data of ModelNet40, because some triangular meshes of ModelNet40 are non-manifold or unclosed. IMMAT merely produced 47.5% MAT data in 13 categories of ShapeNet. In this work, no 3D skeleton is taken as supervision, thus avoiding the limitation of time-consuming data preparation.

2.2 Supervised 3D Shape Reconstruction from a Single-view Image

Single-view image reconstruction only needs an RGB image to obtain a visually realistic 3D model, which is less costly and more user-friendly, thus many research works have been derived to complete this task. Supervised single-view reconstruction methods [7, 8, 11, 14, 25, 25, 26, 28, 30–32, 35, 36, 38, 39, 41, 44] can reconstruct more accurate 3D shapes, but they often require a

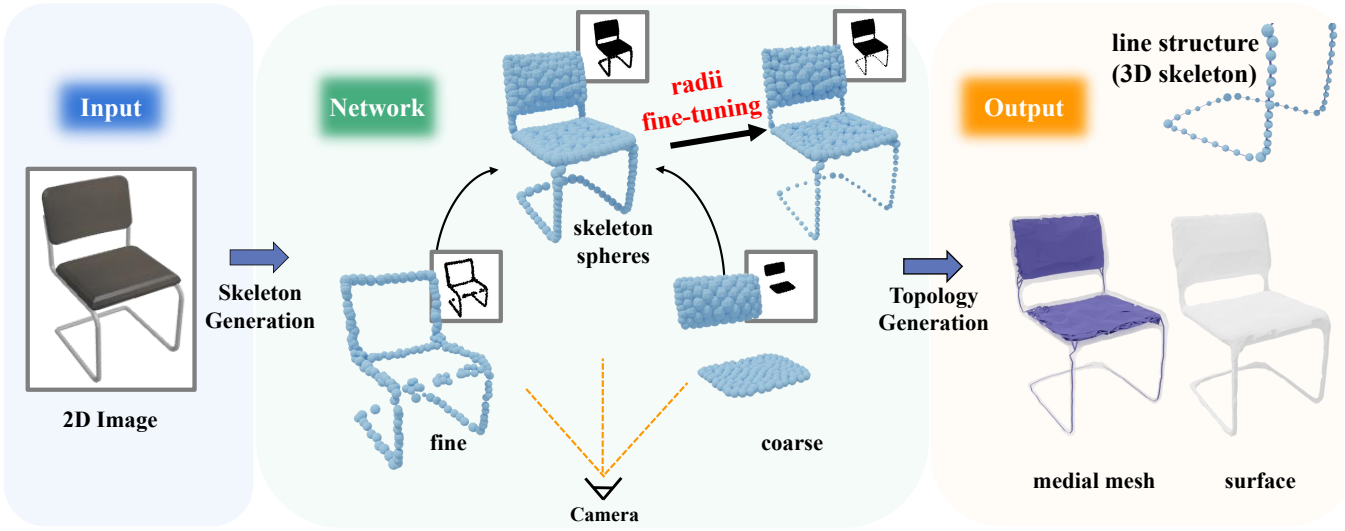

Figure 2: Overview of our S3DS: We generate 3D skeleton spheres by learning coarse and fine semantic structures. We fine-tune the radii to get more accurate volume information. A connectivity generation method constructs the connection relationships of skeleton spheres to form a medial mesh, and then a surface can be reconstructed from its enveloping shape.

large number of real 3D data, which greatly restricts the generalization of the method and costs a lot of time to prepare the data. The reconstructed mesh of Pixel2mesh [38] is deformed from the template spherical mesh, leading to topology constraints of the template, making it difficult to reconstruct objects with holes. Although AtlasNet [12] and TMNet [30] have solved this problem, the reconstruction grid is non-watertight. The implicit methods, such as DISN [41] and OccNet [25], need to calculate the corresponding signed distance function and occupancy value from the ground truth triangular mesh in advance. Skeleton-bridged method [35] uses skeleton points as supervision to improve the reconstruction quality of complex objects. IMMAT [14] further applies MAT to the generation of complex objects.

2.3 Self-supervised Learning on Triangular Mesh

Differentiable rendering for self-supervised learning methods [1, 6, 10, 16–18, 20, 22, 24, 27, 27] are proposed to reconstruct 3D shape by using 2D supervision. NMR [18] proposed an approximate method to integrate rendering into neural networks. Softras [23] uses discrete rasterization and z-buffering as differentiable probabilistic processes to achieve truly differentiable rendering but cannot solve the problems of shadows and topology changes. DIB-R [6] proposed a differential interpolation-based renderer, which computes the gradient analytically.

3 OUR METHOD

3.1 Overview

Given a single-view image of a 3D shape as input, our goal is to obtain the medial spheres and their corresponding connectivity. 3D medial spheres, with radii, are crucial for reconstructing a complete shape, thus essential for differential rendering and self-supervised learning. To this end, we predict the center coordinates and radii of the 3D spheres and then render them into a 2D silhouette in Sec. 3.2. Supervised by the ground-truth silhouette generated from the input image, we expect the projection of the 3D spheres covers as much local shape as possible, which means the 2D rendered circles are close to the authentic 2D medial axis. Due to the symmetric property of the medial axis, as a result, the centers of corresponding 3D medial spheres will be close to ground-truth 3D skeletons.

To overcome the disadvantage of current differentiable methods, which only focus on large patches of the shape, we divide the silhouette into fine and coarse parts, and semantically learn the medial spheres for each part respectively, as described in Sec. 3.3.

In Sec. 3.4, we construct the initial edges of the 3D spheres and fine-tune the radii of medial spheres by optimizing the rendering results of the constructed medial edges. Finally, Sec. 3.5 proposes a simple heuristic method to construct the final medial mesh.

3.2 Skeleton Sphere Reconstruction

We define a medial sphere set $S = (C, R)$, where $C = (X, Y, Z)$ and R are the centers and radii of medial spheres. As shown in Figure 3. Our network contains two image encoders. The fine encoder generates spheres for fine parts of the shape, e.g., chair leg, and the coarse encoder generates spheres for large patches of the shape, e.g., chair back. The fine spheres and the coarse spheres are then rendered to generate the fine silhouette and the coarse silhouette respectively in Sec. 3.3. The two encoders learn their own spheres distributions and do not share weights.

The encoders encode the same input image I into two latent codes T_{fine} and T_{coarse}. Then the latent codes will be fed into two decoders, formed by fully connected layers, and generate two sets

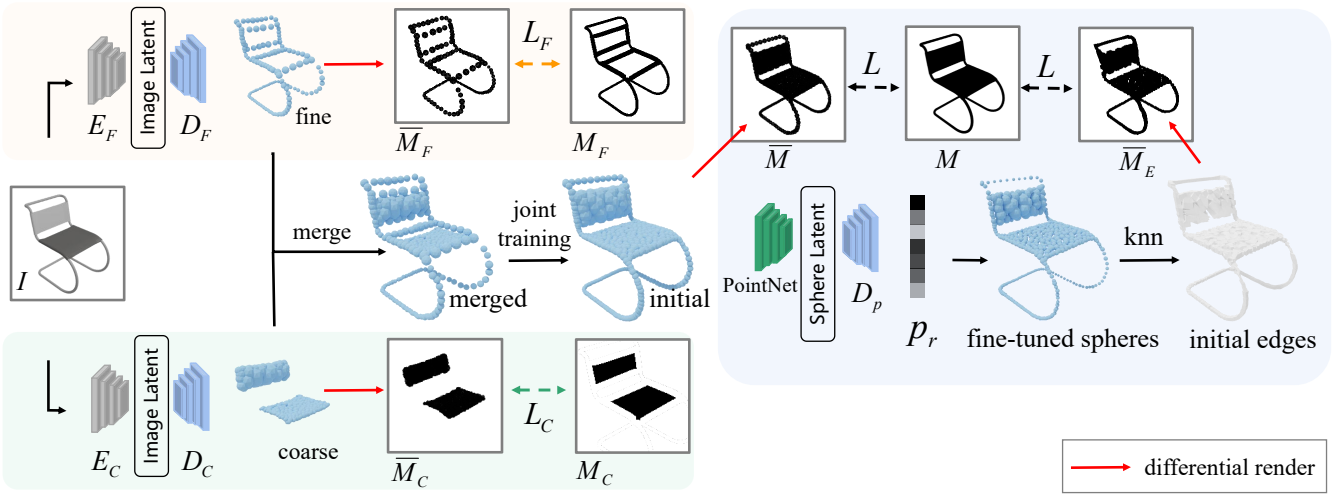

Figure 3: The overall pipeline of semantic sphere learning module (Sec. 3.3). The module uses 2D supervision to reconstruct the skeleton structures with different semantics, and then integrate them through joint training. Finally, the skeleton spheres are fine-tuned with the radius fine-tuning module. We construct initial edges and transfer them to cone mesh to render an edge silhouette.

of spheres respectively by the following formulas:

$$\begin{cases} C = D_c\left(E(I)\right), \\ R = t * \text{sigmoid}\left(D_r\left(E(I)\right)\right), \end{cases} \tag{1}$$

Here E is an image encoder. D_c and D_r are the decoders of spheres' center coordinates and radii. We use $t * \text{sigmoid}(\cdot)$ to limit the radius value in the range $(0, t)$.

Differentiable Rendering on Medial Spheres. A higher resolution 3D sphere mesh contains a larger number of mesh faces, which is time-consuming for a mesh-based differential rendering network. Since the projection of a 3D sphere is a 2D circle, we facilitate the computational cost by rendering 2D circles instead. We use 2D regular octagon mesh with 8 faces, that are orthogonal to the direction of the camera, to approximate the 2D circle. The center vertex of the regular octagon mesh aligns with the center of the corresponding 3D sphere. The distance from the boundary vertex of the octagon mesh to its center is the same as the radius of the sphere.

Once N octagon meshes, which are rotated to orthogonal to camera directions, are generated, we use a differentiable mesh renderer *Rend* to get the rendering silhouette $\bar{M}$ [22], represented by the formula:

$$\bar{M} = Rend\left(Rot_{cam}(X)\right), \tag{2}$$

where $Rot_{cam}(X)$ represents the octagon meshes X rotated according to the camera pose.

3.3 Semantic Sphere Learning

Softras [22] uses the whole silhouette of the input image as supervision to reconstruct 3D shapes. Since the patches take a large portion of pixels in the 2D image, we have observed that the neural network pays more attention to this type of coarse structure while ignoring the fine structure, e.g., thin tubes. In the end, this will cause the missing thin structures in the generated results.

However, the fine structures also reflect the key features of the shapes. Human beings recognize a shape not only by observing the large areas but also by highlighting small regions to improve their perceptions of the shape. Inspired by this observation, we divide the rendered image of the 2D silhouette into two images, one for fine structure and another for coarse structure.

To generate these two types of silhouette images M_F and M_C, we first pixel-divide the silhouette of the target image. Given a $N \times N$ silhouette M, the value of each pixel is $M_{i,j} = 0, 1$, where 1 represents the foreground pixel and 0 represents the background pixel. In order to classify the semantic of each pixel $M_{i,j}$, we take a patch $P_{i,j}$ of size $k \times k$ to calculate the proportion of the mask value of 1 in $P_{i,j}$. The proportion is formulated as $p = \sum(M_{i,j})/(k \times k)$, $p \in [0, 1]$. If $p = 1$, the pixel belongs to the coarse image M_C, otherwise to the fine image M_F.

As shown in Figure 3, we use the two complementary silhouettes M_F and M_C as 2D supervision, to train two encoder-decoder networks for learning the medial spheres in the fine and the coarse structures, namely fine spheres and coarse spheres.

Since the two silhouettes are complementary, theoretically, the reconstructed 3D spheres are also complementary. However, since the two parts are learned independently, the reconstruction results cannot be perfectly merged (Figure 3). To solve this problem, we jointly train the two sub-networks for fine and coarse sphere reconstruction. Specifically, these two sets of spheres are merged and rendered to generate the complete silhouette $\bar{M}$. Then we use the ground truth silhouette M to supervise the training, in order to improve the merged spheres.

Losses. The loss functions are essential for learning accurate and uniformly distributed skeleton spheres. First, we use the IoU loss L_{iou} between the ground truth silhouette and the rendered silhouette, which constrains the rendered circles to be as close to

the canonical 2D medial axis as possible, that is,

$$L_{iou} = 1 - \bar{M} \odot M / \bar{M} \oplus M, \tag{3}$$

where $\odot$ and $\oplus$ are intersection and union operations. When jointly training the whole shape, we multiply the rendered silhouette $\bar{M}$ with the ground truth fine silhouette M_F. Then the 2D IoU loss $L_{attention}$ of M_F is computed to impose additional constraints on the fine structure for better reconstruction results, as shown below:

$$L_{attention} = 1 - (\bar{M} \odot M_F) \odot M_F / (\bar{M} \odot M_F) \oplus M_F. \tag{4}$$

To ensure the consistency of the reconstruction results of the same object from different views, we follow the strategy of Softras[23] and render the reconstruction results of an image onto two views in the training phase. Note that, only one image is used as input in the inference stage. As shown in Eq. 5, L_{2D} represents L_{iou} and $L_{attention}$, X_A is the reconstructed spheres of the image from view A, M_A and M_B are the real silhouettes of view A and view B, $\bar{M}_A$ and $\bar{M}_B$ are the rendered silhouettes of X_A from view A and view B.

$$L_{silhouette} = L_{2D}(\bar{M}_A(X_A), M_A) + L_{2D}(\bar{M}_B(X_A), M_B), \tag{5}$$

Secondly, the 3D sphere loss $L_{nearr} = \sum_{p \in S} |r_p - r_q|$ is introduced to optimize the radius distribution of the spheres, under the assumption that adjacent spheres should share similar radii, q is the nearest sphere of p.

As shown in Figure 3, we use a weighted linear combination of IoU loss and sphere loss to compute L_C, L_F, and L.

Besides, two regularization losses are introduced to generate uniformly distributed medial spheres. The repulsion loss $L_{rep} = \sum_{p \in S} \frac{1}{\|c_p - c_q\|^3}$ uses repulsive force among the neighbor spheres to avoid overlapping. The variance of the surface distance $L_{var} = Var(\|c_p - c_q\| - r_p - r_q), p, q \in S$ between the nearest sphere pairs is also introduced to optimize the distribution of spheres and make the nearest sphere pairs as close as possible. $\|c_p - c_q\|$ is the center distance between sphere p and its nearest sphere q.

3.4 Radius Fine-tuning

Since the learned medial spheres are discrete in 3D space, this will result in un-connected rendering circles in 2D (see $\bar{M}_F$ in Figure 3).

As a result, the network is inclined to increase the radii of the spheres to cover more areas of the GT silhouette, which may generate large errors in thickness.

To address this issue, we propose a radius fine-tuning strategy based on edge rendering. As shown in Figure 3 right blue part, we first learn to scale the radii of the spheres, then constructs initial edges to render an edge silhouette $\bar{M}_E$. The scale factors are optimized by computing the 2D IoU loss L between the $\bar{M}_E$ and GT silhouette M.

More specifically, we use PointNet [33] to encode the initial spheres to a latent code and then decode the code to N scale factors $Pr \in (0, 1)$ to scale the initial spheres to fine-tuned spheres. After fine-tuning, we construct initial edges by assuming that each sphere connects to 2 nearby spheres. To render 3D edges, similar to how we render 3D spheres, we use the 2D trapezoid mesh *Cone* as an approximation of the projection of medial cone, namely as *cone mesh* in Figure 4 right. The height of the cone mesh is the length of the medial edge (distance from one sphere center to another), and the value of the upper base (lower base) is equal to the diameter of the upper sphere (lower sphere). Obviously, the sphere radius affects the thickness of the cone. And the edge silhouette $\bar{M}_E$ is rendered as a union of $2N$ cone meshes. In this way, the model will learn spheres with more accurate radii and render edges to fill the gap between the discrete rendered circles. In the implementation, $Cone(i, j) = (h, r_i, r_j)$, h is the edge length, and r_i, r_j are radii of spheres.

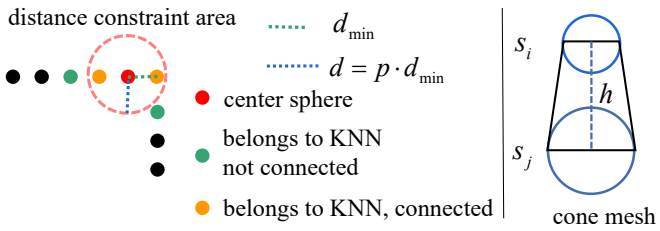

Figure 4: Connectivity generation and cone mesh visualized in 2D.

3.5 Connectivity Generation

We propose a heuristic method to generate reasonable medial edges and medial faces of the medial spheres for MAT. Up till now, the learned medial spheres are already close to the 3D shape in spatial distribution, all we need is just a simple connection. Different from Sec. 3.4 where we only generate 2 edges for each sphere, here we use a larger K to generate more edges ($K \geq 2$). To reduce the bad connections, we compute the distances between the center sphere and its adjacent spheres to determine whether they can be connected or not. We depict a simple diagram (see Figure 4 left) to demonstrate the edge generation in this step.

The key idea is to add filtering on the number and the distance of edges. Specifically, 1) We set different K for the two types of semantic spheres. Fine spheres usually require fewer edges, which can form a curved structure, e.g., chair legs, while coarse spheres need more edges so that rendered medial cones can cover a larger portion of the silhouette.

2) We only generate the edges for each sphere in a limited scope. We first compute the nearest distance d_{min} of the center sphere to its neighbors and set a distance threshold $d = p \cdot d_{min}, p > 1$.

One edge will be selected only if the length of the edge is less than d. Using a fixed threshold d, for example, the constructed edges are either too few or too many for areas with varying densities. Hence, associating d with d_{min} gives us the ability to be more adaptive to densities. As shown in Figure 4, the center sphere has at most 4 adjacent spheres (green and orange), but we only select the spheres (orange) within the distance d to generate edges.

This strategy takes both the number and the distance of edges into consideration for constructing more canonical connectivity. Similarly, the medial faces are generated based on triangle edges. Combining all three elements (spheres, medial edges, medial faces) gives us a complete Medial Axis Transform (MAT) structure.

Table 1: Quantitative comparisons with self-supervised (Soft-SK) and supervised (SkeletonBridge, IMMAT) methods on 3D skeletons reconstruction from images. The best results are boldfaced.

Method	Q-MAT(CD)		Q-MAT(R)		DPC(CD)		Q-MAT(CD)		Q-MAT(R)	
	Soft-SK	S3DS	Soft-SK	S3DS	SkeletonBridge	S3DS	IMMAT	S3DS	IMMAT	S3DS
Plane	0.94	**0.72**	**1.39**	1.43	0.40	**0.34**	**0.27**	0.32	**0.47**	1.31
Bench	1.29	**1.12**	1.43	**0.87**	0.52	**0.50**	0.41	**0.40**	**0.52**	0.82
Dresser	**1.51**	1.54	**3.29**	3.84	**0.63**	0.64	**0.50**	0.62	**1.70**	3.30
Car	1.18	**1.13**	2.81	**2.78**	**0.50**	0.52	**0.41**	0.52	**1.53**	2.87
Chair	1.33	**1.22**	1.50	**1.17**	0.56	**0.45**	0.50	**0.44**	**0.79**	1.12
Display	1.43	**1.38**	2.16	**1.76**	**0.71**	0.74	**0.52**	0.53	**1.16**	1.54
Lamp	1.38	**1.21**	2.06	**1.38**	0.64	**0.63**	**0.43**	0.49	**1.03**	1.34
Speaker	**1.39**	1.45	**3.21**	3.54	0.78	**0.53**	**0.65**	0.66	**2.37**	2.43
Rifle	0.74	**0.49**	1.01	**0.66**	0.43	**0.33**	**0.22**	0.25	**0.51**	0.65
Sofa	**1.47**	1.51	**2.16**	2.29	0.60	**0.55**	**0.53**	0.66	**1.68**	2.66
Table	1.48	**1.38**	1.89	**1.32**	0.56	**0.49**	**0.54**	0.55	**0.80**	1.24
Phone	**1.29**	1.30	**1.47**	1.48	0.44	**0.38**	**0.36**	0.42	**1.01**	1.34
Vessel	1.09	**0.92**	1.92	**1.63**	0.53	**0.41**	**0.38**	0.41	**1.15**	1.62
Mean	1.27	**1.13**	2.02	**1.48**	0.54	**0.47**	**0.44**	0.47	**0.86**	1.40

4 EXPERIMENTS

4.1 Experiments on Synthetic Datasets

Implement Details. The regular polygon circular mesh contains 9 vertices and 8 faces, while the cone mesh contains 8 vertices and 12 faces. The size of the input image and the rendered silhouette are both 224x224. Following Softras [22], two images from different views of the same object are used for training, and the camera poses of the two views are used. The inference stage uses one image and does not need the camera pose. The renderer is Kaolin [9]. The two encoders of the network are pre-trained ResNet18 [13], and the decoder consists of 3 fully connected layers. The batch size is 64 and the initial learning rate is 1e-4.

The numbers of both fine and coarse spheres are 200, and the maximum radius is 0.2. In the radius fine-tuning stage, we only optimize the radii of the spheres.

We also designed a baseline Soft-SK (a variant of Softras [22] that reconstructs the 3D skeletons). On the basis of the Softras model, we directly use N spheres to replace the spherical mesh and predict the medial spheres.

Dataset. We conduct experiments on the synthetic dataset ShapeNet and the dataset of real image pix3d [34]. We follow NMR [18] to use blender [3] to render color images of 13 categories of ShapeNet. ShapeNet is divided into widely used training and test sets [23]. We render images of each CAD model from 24 azimuth angles with a fixed elevation angle of 30°. For the training set, we render 24 additional images at 0°.

Comparison on 3D skeletons. We evaluate the reconstruction performances of 3D skeletons on the ShapeNet and conduct quantitative and qualitative comparisons with the self-supervised methods Soft-SK and the supervised methods SkeletonBridge [35] and IMMAT [14].

For comparison with Soft-SK, we use the GT MAT data provided by MAT-Net [14], which computes the MATs of 256 spheres by Q-MAT [19]. Since Q-MAT couldn't compute MATs from the non-manifold meshes in ShapeNet, we only select the samples that intersect with the image test set for comparison. We adopt Chamfer Distance (CD) on Radii Distance losses [14] as comparison metrics on medial spheres. As shown in Table 1, S3DS outperforms the Soft-SK on both the average metrics of all categories of samples, indicating that our approach has brought significant improvement in general. The qualitative results shown in Figure 5 demonstrate that our method could reconstruct better 3D skeletons than the baseline, even for the complex shapes. The learned 3D skeletons with our method are a string of spheres in the tubular structure or a layer in the coarse structure. At the same time, the centers of the spheres are located on the medial axis. These results illustrate that by only using a 2D image, the 3D skeletons can be reconstructed well. Although the baseline can reconstruct most of the skeletons of the shape, it failed to reconstruct the fine structures in some samples, i.e., the frame of the lamp in the fourth column. Besides, our method couldn't compute the 3D skeletons as well as Soft-SK in some categories. To analyze the results of these categories, we statistics the number of each category and found that the baseline outperforms our method for categories with fewer samples. For other categories, such as sofa, the shapes are almost homomorphic to a sphere. There are fewer spheres in the fine structure of these shapes, and the spheres have little impact on the quantification.

For fair comparisons, we retrain SkeletonBridge and IMMAT on the corresponding dataset, that is 96.0% and 47.5% objects of ShapeNet, and train and test our model on the same dataset as these two methods. For SkeletonBridge, we use the CD loss between the reconstruction results and the meso-skeleton as the metric.

In Table 1, although SkeletonBridge employs meso-skeleton as supervisory information, S3DS performs better than SkeletonBridge quantitatively. Accordingly, as shown in Figure 6, the reconstructed

Figure 5: Qualitative comparison with the competitive self-supervised 3D skeletons reconstruction methods from images.

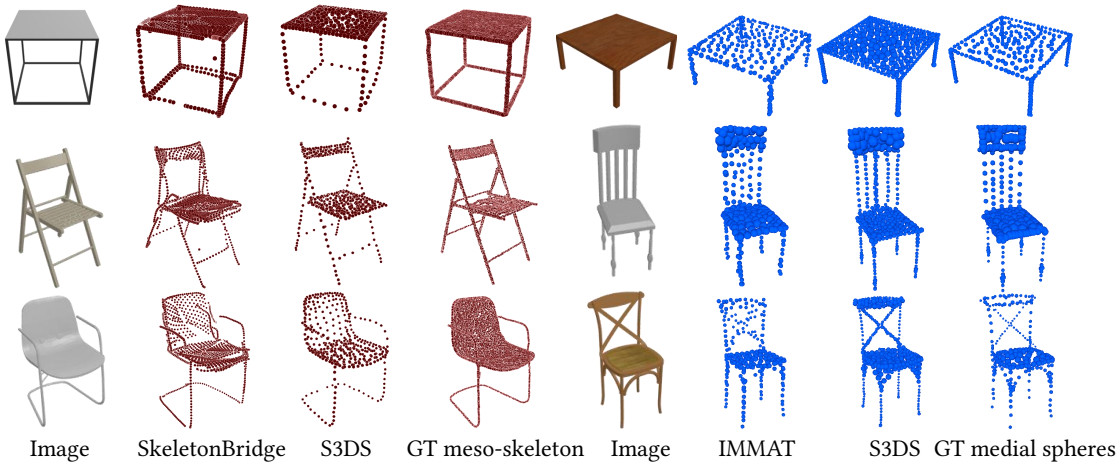

Figure 6: Qualitative comparisons with supervised methods on 3D skeletons reconstruction.

centers by S3DS are also visually closer to the ground truth meso-skeleton. This is because the skeletons reconstructed by SkeletonBridge are only used to construct rough geometry, and the ultimate goal of SkeletonBridge is not to learn precise skeletons. The advanced medial spheres reconstruction method IMMAT is quantitatively superior to S3DS because IMMAT directly optimizes CD loss and radius loss during training supervised by the ground truth medial spheres computed with Q-MAT [19], directly leading to better quantitative results of IMMAT. In contrast, S3DS does not optimize these two loss functions. However, as shown in Figure 6, S3DS can also learn visually approximate results of the medial spheres than IMMAT, especially in center distribution. In summary, S3DS did not include any guidance on meso-skeleton and medial spheres in design, but the reconstruction results have the features of 3D skeletons while quantitatively surpassing or approaching the relevant methods.

Comparison on Connectivity Generation. On the connectivity generation of medial axis transform, We compare our method with traditional methods, including Delaunay Triangulation (deleting overlong edges), Ball Pivoting [2], and K Nearest Neighbor (KNN) (directly connecting K nearest neighboring spheres for each sphere to form medial faces).

As shown in Figure 7, Delaunay Triangulation (constrain edge length) or KNN (constrain edge number) can reconstruct the general connectivity between spheres. However, they are not capable to construct fine tubular structures in sharp parts. Our connectivity generation strategy comprehensively considers the edge number and distance threshold constraints of spheres with the fine and coarse semantics, so that we can simultaneously construct a complete plane structure and a fine tubular structure.

Figure 7: Comparison on connectivity generation methods.

Testing on Real Images. To further validate the effectiveness of S3DS, experiments on the real image dataset Pix3d [34] are

Figure 8: Results on real images.

conducted. Reconstruction results in Figure 8 show that our method could not only learn the overall shape of the object but also the complex details, such as the holes which benefit from the learning of the two semantic structures. The experimental process on real image data is described in the appendix.

Surface Mesh Reconstruction. Following [14], the surface mesh can be reconstructed from the generated MAT. We compared our method with classically related surface mesh reconstruction works P2M [38] and Softras [23] on the category of chairs. We retrain the methods on the same datasets.

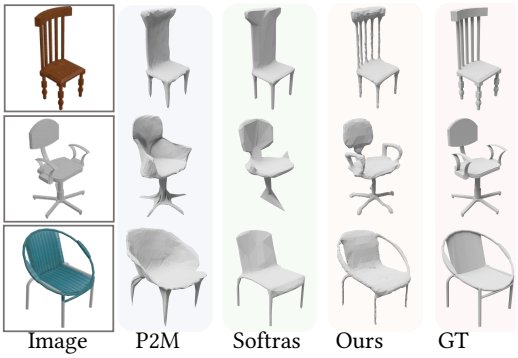

Figure 9: Qualitative results on mesh reconstruction.

Table 2: Quantitative comparison of reconstruction.

3D IoU ↑	w/o holes	w/ holes	mean	supervision
Softras	0.556	0.496	0.523	2D image (2 views)
P2M	**0.607**	0.495	**0.546**	3D mesh
Ours	0.532	**0.536**	0.534	2D image (2 views)

Figure 9 shows that our method can generate complex shapes and has an advantage in reconstructing tubular structures. We divide the chairs into two subsets according to whether they have holes for quantitative comparison. The chair without holes is homeomorphic to the sphere, making it easy to deform from the spherical mesh. The chair without holes is more complex and difficult to reconstruct, as shown in Figure 9. We use the IoU of the reconstructed mesh and the ground truth mesh, denoted as 3D IoU in Table 2, to measure the reconstruction results. As shown in Table 2, in terms of mean error, our method is better than Softras, which is also using the 2D image as supervision, and worse than P2M, because P2M uses 3D mesh as the target constraint. Results in Figure 9 show that P2M could not work well on complex shapes, while our method could.

In addition, our quantitative difference between the two types is very small, indicating that our method has more stable performance.

4.2 Ablation Study

Ablation of joint training: The fine spheres and the coarse spheres obtained through separate training are often inconsistent in spatial distribution and cannot be directly merged.
Ablation of semantic sphere learning: Without semantic sphere learning, some tubular structures are missing.
Ablation of radius fine-tuning: The spheres will have large radii when without fine-tuning.

Figure 10 shows our full model can generate complete medial spheres with accurate radii. Table 3 shows that the full model has the best quantitative results.

Table 3: Quantitative comparisons of different strategies.

Model	CD (Sphere)	R (Sphere)
w/o jointly training	1.27	1.67
w/o semantic	1.25	1.74
w/o fine-tuning	**1.13**	1.93
full	**1.13**	**1.48**

4.3 Limitation and Discussion

The synthesized dataset used in S3DS has a fixed perspective, making it sensitive and lacking robustness. Although we have achieved visually remarkable 3D skeletons, the reconstruction performance for concave shapes is currently limited due to depth-related challenges. S3DS demonstrates superior performance compared to other methods in shape reconstruction, particularly when dealing with clearly defined skeletal structures. Additionally, it proves effective in reconstructing shapes (dresser, car) that lack obvious skeletal structures. Experimental analysis reveals that S3DS is suitable for large-scale datasets encompassing a diverse range of shape classes.

5 CONCLUSION

In this paper, we propose the first self-supervised method of reconstructing the 3D skeleton of a shape from its single-view image, namely S3DS, by using medial axis transform as the underlying representation. With the semantic sphere learning module and the radius fine-tuning strategy, the precise distribution of medial spheres as well as their radii is learned. Then a heuristic strategy is used for reconstructing the connectivity of MAT from the learned medial spheres, generating a complete medial mesh. The experimental results show that it is effective in learning 3D skeletons by fitting 2D images through differential rendering, avoiding the problem of the time-consuming preparation of 3D skeleton data for supervision. Compared with the baseline that directly uses an overall silhouette, our method is more accurate in the learning of skeletal spheres and has a better visual effect. Our method also achieves competitive results compared to the supervision methods.

REFERENCES

[1] Kalyan Vasudev Alwala, Abhinav Gupta, and Shubham Tulsiani. 2022. Pre-train, Self-train, Distill: A simple recipe for Supersizing 3D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3773–3782.

[2] Fausto Bernardini, Joshua Mittleman, Holly Rushmeier, Claudio Silva, and Gabriel Taubin. 1999. The ball-pivoting algorithm for surface reconstruction. *IEEE transactions on visualization and computer graphics* 5, 4 (1999), 349–359.

[3] Blender Online Community. 2022. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam. http://www.blender.org

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

[5] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, and Hao Su. 2015. ShapeNet: An Information-Rich 3D Model Repository. *Computer Science* (2015).

[6] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. 2019. Learning to predict 3d objects with an interpolation-based differentiable renderer. *Advances in Neural Information Processing Systems* 32 (2019).

[7] Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. 2020. Bsp-net: Generating compact meshes via binary space partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 45–54.

[8] Christopher B. Choy, Danfei Xu, Jun Young Gwak, Kevin Chen, and Silvio Savarese. 2016. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. In *European Conference on Computer Vision*.

[9] Clement Fuji Tsang, Maria Shugrina, Jean Francois Lafleche, Towaki Takikawa, Jiehan Wang, Charles Loop, Wenzheng Chen, Krishna Murthy Jatavallabhula, Edward Smith, Artem Rozantsev, Or Perel, Tianchang Shen, Jun Gao, Sanja Fidler, Gavriel State, Jason Gorski, Tommy Xiang, Jianing Li, Michael Li, and Rev Lebaredian. 2022. Kaolin: A Pytorch Library for Accelerating 3D Deep Learning Research. https://github.com/NVIDIAGameWorks/kaolin.

[10] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. 2022. Get3d: A generative model of high quality 3d textured shapes learned from images. *arXiv preprint arXiv:2209.11163* (2022).

[11] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas A. Funkhouser. 2019. Deep Structured Implicit Functions. *CoRR* abs/1912.06126 (2019). http://arxiv.org/abs/1912.06126

[12] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. 2018. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Jianwei Hu, Gang Chen, Baorong Yang, Ningna Wang, Xiaohu Guo, and Bin Wang. 2022. IMMAT: Mesh Reconstruction from Single View Images by Medial Axis Transform Prediction. *Computer-Aided Design* 150 (2022), 103304. https://doi.org/10.1016/j.cad.2022.103304

[15] Jianwei Hu, Bin Wang, Lihui Qian, Yiling Pan, Xiaohu Guo, Lingjie Liu, and Wenping Wang. 2019. MAT-Net: Medial Axis Transform Network for 3D Object Recognition.. In *IJCAI*. 774–781.

[16] Tao Hu, Liwei Wang, Xiaogang Xu, Shu Liu, and Jiaya Jia. 2021. Self-Supervised 3D Mesh Reconstruction from Single Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6002–6011.

[17] Eldar Insafutdinov and Alexey Dosovitskiy. 2018. Unsupervised learning of shape and pose with differentiable point clouds. *Advances in neural information processing systems* 31 (2018).

[18] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3907–3916.

[19] Pan Li, Bin Wang, Feng Sun, Xiaohu Guo, Caiming Zhang, and Wenping Wang. 2015. Q-MAT:Computing Medial Axis Transform By Quadratic Error Minimization. *ACM Transactions on Graphics* 35, 1 (2015), 1–16.

[20] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. 2020. Self-supervised single-view 3d reconstruction via semantic consistency. In *European Conference on Computer Vision*. Springer, 677–693.

[21] Cheng Lin, Changjian Li, Yuan Liu, Nenglun Chen, Yi-King Choi, and Wenping Wang. 2020. Point2Skeleton: Learning Skeletal Representations from Point Clouds. *CoRR* abs/2012.00230 (2020). arXiv:2012.00230 [cs.CV]

[22] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. 2019. Soft Rasterizer: A Differentiable Renderer for Image-Based 3D Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[23] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. 2019. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7708–7717.

[24] Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. 2019. Learning to infer implicit surfaces without 3d supervision. *Advances in Neural Information Processing Systems* 32 (2019).

[25] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] Mateusz Michalkiewicz, Jhony Kaesemodel Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. 2019. Implicit Surface Representations As Layers in Neural Networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

[27] KL Navaneet, Ansu Mathew, Shashank Kashyap, Wei-Chih Hung, Varun Jampani, and R Venkatesh Babu. 2020. From image collections to point clouds with self-supervised shape and pose networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1132–1140.

[28] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. 2020. Total3DUnderstanding: Joint Layout, Object Pose and Mesh Reconstruction for Indoor Scenes from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 55–64.

[29] Yinyu Nie, Yiqun Lin, Xiaoguang Han, Shihui Guo, Jian Chang, Shuguang Cui, Jian Zhang, et al. 2020. Skeleton-bridged point completion: From global inference to local adjustment. *Advances in Neural Information Processing Systems* 33 (2020), 16119–16130.

[30] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. 2020. Deep Mesh Reconstruction From Single RGB Images via Topology Modification Networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

[31] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 165–174.

[32] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[33] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 652–660.

[34] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. 2018. Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Jiapeng Tang, Xiaoguang Han, Junyi Pan, Kui Jia, and Xin Tong. 2019. 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*. 4541–4550.

[36] Jiapeng Tang, Xiaoguang Han, Mingkui Tan, Xin Tong, and Kui Jia. 2021. Skeletonnet: A topology-preserving solution for learning mesh reconstruction of object surfaces from rgb images. *IEEE transactions on pattern analysis and machine intelligence* (2021).

[37] Ningna Wang, Bin Wang, Wenping Wang, and Xiaohu Guo. 2022. Computing Medial Axis Transform with Feature Preservation via Restricted Power Diagram. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–18.

[38] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. 2018. Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[39] Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 2019. 3DN: 3D Deformation Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[40] Shihao Wu, , Hui Huang, Minglun Gong, Matthias Zwicker, and Daniel Cohen-Or. 2015. Deep Points Consolidation. *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia 2015)* 34, 6 (2015), 176:1–176:13.

[41] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 2019. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *Advances in Neural Information Processing Systems*. 492–502.

[42] Yajie Yan, David Letscher, and Tao Ju. 2018. Voxel cores: Efficient, robust, and provably good approximation of 3d medial axes. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.

[43] Baorong Yang, Junfeng Yao, Bin Wang, Jianwei Hu, Yiling Pan, Tianxiang Pan, Wenping Wang, and Xiaohu Guo. 2020. P2MAT-NET: Learning medial axis transform from sparse point clouds. *Computer Aided Geometric Design* 80 (2020), 101874.

[44] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. 2017. FoldingNet: Interpretable Unsupervised Learning on 3D Point Clouds. *CoRR* abs/1712.07262 (2017). arXiv:1712.07262 http://arxiv.org/abs/1712.07262

A APPENDIX

A.1 Ablation Study

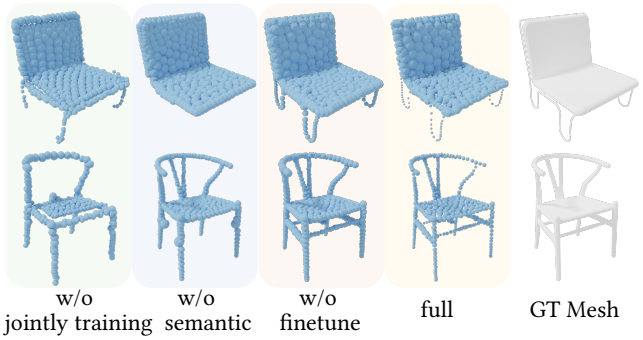

Figure 10: Visual effects of different ablations.

Ablation on losses. To further validate the effectiveness of the losses we used, we conducted ablations studies on the three categories (chair, car, table) by removing the corresponding losses in sequence. As shown in Table 4, removing the regularization losses L_{rep} and L_{var} will increase both the CD loss and the R loss between the predicted spheres and the ground truth spheres. The lack of radius loss has little impact on the CD loss but has a serious impact on the R loss which measures the radius consistency.

Table 4: Quantitative comparisons of different losses.

Model	CD (Sphere)	R (Sphere)
w/o L_{rep}, L_{var}	1.32	1.89
w/o L_{nearr}	1.29	1.95
full	**1.21**	**1.75**

A.2 More visual samples

In Figure 11, we show the results of the predicted medial spheres (3D skeleton with radii) using our method in more categories. It can be seen that our method can generate better skeleton structures in various categories, especially in regions with thin tube and frame structures. For example, our method predicts proper spheres on thin arms and legs for bench models (columns 4-7), and the correct square shape of the lamp base (column 8) instead of the rounded shape predicted from Soft-SK (see Sec. 4.1).

A.3 Analysis of mesh reconstruction

We further analyze the mesh reconstruction results. As shown in Figure 12, methods based on spherical mesh deformation, e.g., P2M[38] and Softras[22], are more suitable for learning shapes homeomorphic to the shape template. The upper row shows that these methods perform better in shapes without holes. This is because they use the Laplacian smoothing [22] constraint on the mesh to obtain a smoother surface. Moreover, P2M[38] uses the ground truth 3D mesh as supervision, thus can learn concave shapes, e.g., the bases of the chairs. Softras [22] introduced a novel formulation that views rendering as an aggregation function that fuses the probabilistic contributions of all mesh triangles with respect to the rendered pixels to flow gradients to the occluded and far-range vertices.

However, these methods are hard to reconstruct the shapes with holes due to the limitation of the template mesh. Our method, on the contrary, performs much better on models with holes, e.g., the bottom row in Figure 12. In general, our method, which requires only the input images as 2D supervision, is more stable and has more advantages in reconstructing complex shapes.

A.4 Connectivity Generation

Algorithm 1 shows the pseudo-code for generating connectivity given medial spheres $S = (C, R)$, where C is the center coordinates and R the radii. The inputs are predicted medial spheres S and three hyper-parameters K_{fine}, K_{coarse} and p, where K_{fine} and K_{coarse} are maximum numbers of neighboring edges for fine spheres and coarse spheres respectively, and p is a parameter larger than 1 which is the distance threshold for generating the medial edges. (see Sec. 3.5 for more details.)

Although we construct a face by connecting three edges, there are still some holes during reconstruction, so we use the same hole-filling method as Point2Skeleton [21] to solve this problem.

Algorithm 1 Connectivity Generation on Medial Spheres

Input:
Predicted medial spheres: $S = (C, R)$;
K_{fine}; K_{coarse}; p
Output: Edges: E;
Compute neighbor distances and indexes.
$dis_c, idx_c = \text{KNN}(C)$
Select nearest neighbor.
$e_c = idx_c[0]$
Select neighbors by center distances.
$E_c = []$
foreach c in C **do**
if c is fine centers
$K = K_{fine}$
else
$K = K_{coarse}$
$min\_dis = dis_c[0]$
for $i = 1, ..., K$
if $idx_c[i] < p * min\_dis$
add edge to E_c
$E = e_c + E_c$

A.5 The details of comparisons with supervised methods

The retrained SkeletonBridge[35] takes real meso-skeletons as supervision and trains on the images from S3DS. The results of the SkeletonBridge, S3DS, and the ground truth meso-skeletons contain 2600, 400, and 7500 skeletal points respectively. To conduct the quantitative comparison, we randomly sample 400 points from SkeletonBridge and ground truth meso-skeletons.

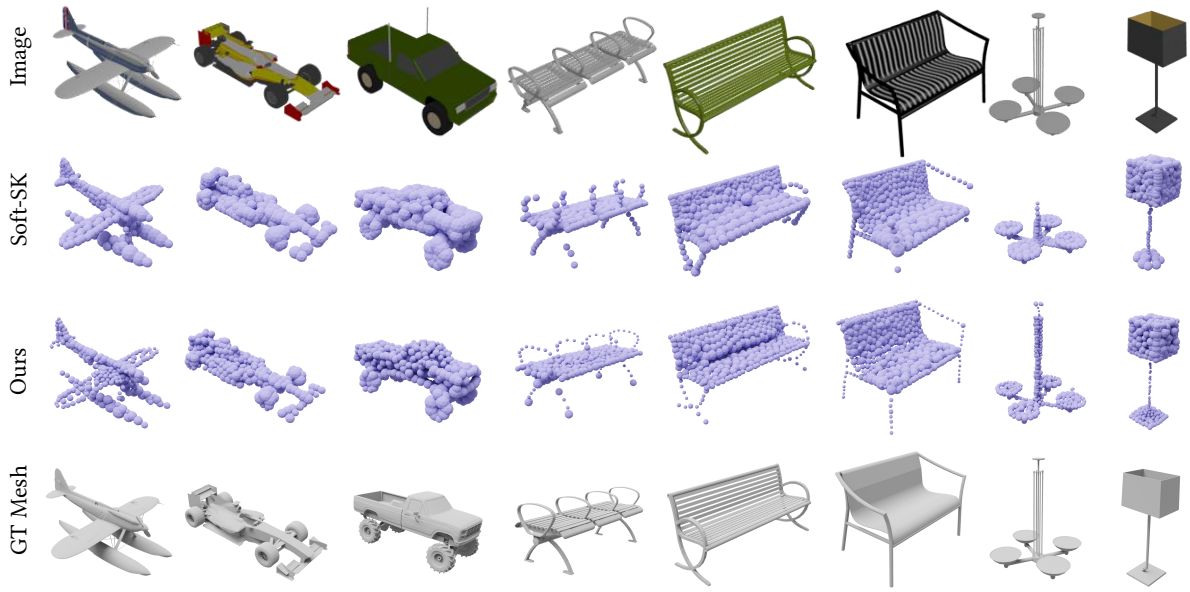

Figure 11: 3D skeleton reconstructions in more categories.

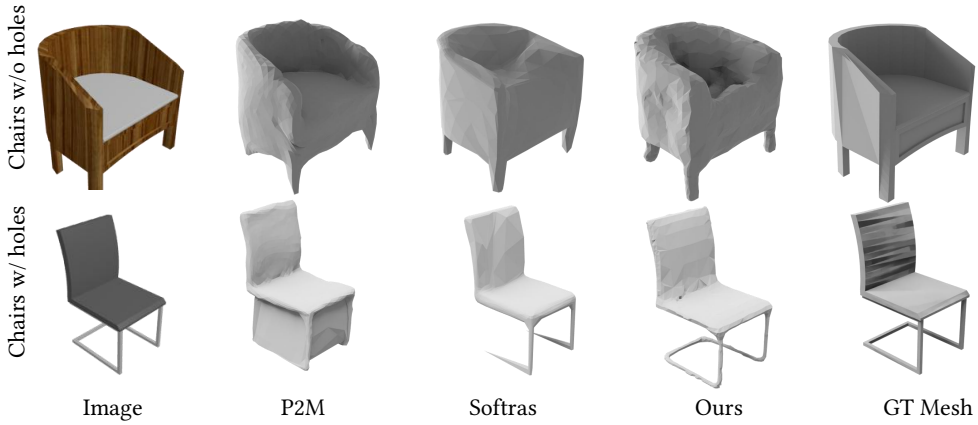

Figure 12: Qualitative results on mesh reconstruction.

We compare S3DS with the results from the refinement stage of the IMMAT[14] because it only uses the medial spheres as supervisory information. The reconstruction and ground truth of IMMAT are 256 spheres, we randomly sample 256 spheres from the results of the S3DS to conduct a quantitative comparison.

A.6 Implementation in real image

Pix3d provides real-world images with complex backgrounds and manually segmented masks. S3DS is trained on a synthetic dataset, and there is a significant difference in distribution between the synthesized image and the real image. To solve this problem, similar to SkeletonBridge, we use a mask to segment the object, move it to the center of the image, and finally resize the image to a size of 224x224. From this, we input it into S3DS to obtain the skeletons. Although the distribution of real images is more complex in viewpoint and scale, the results appear to be consistent with those of synthesized images, which verifies the effectiveness of our method.

A.7 Details of losses

$$\begin{cases} L_C = L_{iou} + \lambda_1 L_{rep} + \lambda_2 L_{var} + \lambda_3 L_{nearr}, \\ L_F = L_{iou} + \lambda_1 L_{rep} + \lambda_2 L_{var} + \lambda_3 L_{nearr}, \end{cases} \tag{6}$$

$$L = L_{iou} + \lambda_0 L_{attention} + \lambda_1 L_{rep} + \lambda_2 L_{var} + \lambda_3 L_{nearr}, \tag{7}$$

In Semantic Sphere Learning (Sec. 3.3) and Radius Fine-tuning (Sec. 3.4) stages, we use loss functions to promote the medial spheres prediction. L_C and L_F are used for the learning of the coarse spheres and the fine spheres respectively. L is applied to all spheres and used in joint training and Radius Fine-tuning. In the experiments, we use $\lambda_0 = 1.0, \lambda_1 = 1e^{-7}, \lambda_2 = 0.1, \lambda_3 = 0.2$. And in radii fine-tuning, the repulsion loss L_{rep} (see Sec. 3.3) and the variance loss L_{var} (see Sec. 3.3) are not used, i.e., $\lambda_1 = 0, \lambda_2 = 0$.