SculptFormer: Transformer Boosted 3D Mesh Reconstruction from 2D Images

Evan Kim (evan_kim@mit.edu) & Shrika Eddula (shrika@mit.edu)
Paper ID *****

Abstract

Reconstructing detailed 3D object shapes from single 2D images is a challenging computer vision task with many important applications, such as creating immersive augmented reality (AR) experiences, enabling intelligent robotic interactions, and generating realistic 3D assets for multimedia. While recent deep learning approaches have made progress, faithfully recovering intricate local geometric details like sharp edges and thin structures, while simultaneously preserving coherent global 3D structures, remains an open challenge. In this work, we propose SculptFormer, a transformer-boosted framework for multi-scale 3D mesh reconstruction from single-view inputs. Inspired by the coarse-to-fine approach of Pixel2Mesh, our architecture enhances the deformation process with transformer components at the global, intermediate, and local levels. Specifically, a global transformer attends to coarse, holistic shape features to control the overall 3D structure prediction while intermediate and local graph-based transformer blocks progressively refines detailed local geometry by attending to lower point features as the 3D mesh is upsampled. Through evaluations on 3D objects taken from 13 object categories in the ShapeNetCore dataset, we find that our approach successfully generates more accurate 3D reconstructions compared to Pixel2Mesh.

1. Introduction

Reconstructing 3D models of objects from 2D images has many downstream applications such as creating realistic and immersive AR/VR experiences and enabling virtual object placement and interaction. 3D shape reconstruction can also aid in object recognition, grasping, and manipulation tasks for robotic systems, enabling more efficient and accurate interactions with the physical world.

1.1. Related Work

Current approaches for single-view 3D shape reconstruction from 2D images can be broadly categorized into voxel-based, mesh-based, and point cloud-based methods. Voxel-based techniques represent the 3D shape as a voxel grid and employ convolutional neural networks (CNNs) or other deep learning models to predict the occupancy value of each voxel given the input 2D image, as explored in works such as 3D-R2N2 [3] and OGN [9]. Alternatively, mesh-based methods directly predict the 3D mesh representation comprising vertices and faces that form the object's surface, with approaches like Pixel2Mesh [10] and AtlasNet [4] being notable examples. Point cloud-based methods predict an unstructured set of 3D points representing the object's geometry from the 2D input, such as PSGN [5].

However, these 3D representations also face significant limitations. Voxel-based approaches can produce high-resolution 3D shapes but are computationally inefficient, especially for large voxel grids, and often exhibit discretization artifacts manifesting as blocky surfaces. Mesh-based predictions are more efficient but can struggle to generate topologically-correct meshes, especially for geometrically complex shapes. Point cloud outputs lack explicit surface information and may suffer from non-uniform point distributions.

Factors such as occlusions, varying viewpoints, cluttered backgrounds, and illumination conditions further exacerbate the complexity of the task. Many current techniques rely heavily on strong priors from object categories, hindering their ability to generalize well to novel object types. Moreover, the lack of coherence and temporal stability in predictions poses challenges for applications requiring consistent reconstructions. Developing an approach capable of robustly reconstructing accurate, high-resolution 3D shapes across diverse real-world settings while preserving fine geometric details remains an open research problem. Novel neural architectures and modeling techniques are necessary to fully unlock the potential of single-view 3D shape reconstruction from limited 2D data.

1.2. Our Method

While recent years have seen significant progress in single-view 3D shape reconstruction, a major ongoing challenge involves simultaneously capturing accurate holistic shape information as well as intricate local geometric details from just a single 2D image [1, 5]. Many existing methods excel at reconstructing the overall coarse 3D structure of an object but fail to faithfully recover fine-grained geometry like sharp edges, thin structures, complex concavities, and precise surface details [2]. Conversely, techniques that aim to generate highly-detailed 3D geometry often struggle with maintaining global coherence and producing plausible holistic 3D shapes [7].

This limitation arises from the inherent difficulty in effectively leveraging the limited visual cues present in a single 2D observation to infer precise 3D shape information at both macro and micro scales. Additionally, existing 3D representation formats like voxel grids [3], point clouds [5], and mesh surfaces [10] have inherent tradeoffs in balancing reconstruction quality, memory efficiency, and geometric expressiveness. Recently, transformer-based architectures [6] have shown promise in integrating local and global information for coherent 3D shape generation via self-attention mechanisms that can capture long-range reconstruction features while also focusing on fine details.

Developing architectures that can seamlessly fuse 3D shape priors at multiple levels of detail to produce coherent, high-fidelity 3D reconstructions remains a challenge. To address this, we propose a novel framework that combines the strengths of mesh-based representations and transformers. Our architecture uses a transformer encoder to extract rich contextual features from the input 2D image, while the transformer decoder generates the 3D voxel representation in an autoregressive manner. Crucially, our decoder employs a hybrid self-attention mechanism that attends to both global, holistic shape information as well as local, fine-grained geometric details. This allows our model to simultaneously keep track of varying levels of overall 3D structure as well as intricate local geometry, overcoming a previous inability to maintain shape coherence at multiple levels of detail at once, and outperforming prior voxel and mesh-based methods on this challenging 3D reconstruction task.

2. Methods

2.1. Base Architecture

In this paper, we will be building off of the existing Pixel2Mesh architecture. Pixel2Mesh [10] is a graph-based deep learning framework designed to generate 3D mesh models directly from a single 2D image input. It employs two main components that work in parallel, the first being VGG-16, which serves to extract features from the input 2D image, and the second being a graph convolutional neural network (GCN) that deforms an initial ellipsoid mesh towards the target 3D shape in a coarse-to-fine manner, initially starting with fewer vertices and higher-level input features. The GCN operates on the mesh vertices and edges, capturing local geodesic information to progressively refines the mesh through the addition of new vertices to increase the representational power of the mesh and successive deformation stages guided by subsequent, lower-level 2D image features.

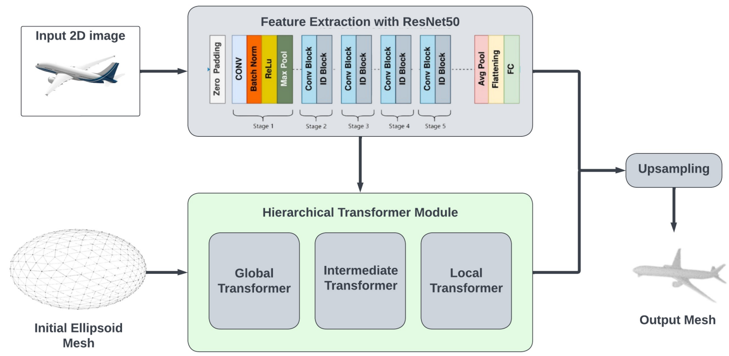

Figure 1. SculptFormer architecture with hierarchical transformer module

We propose SculptFormer, a new framework that extends Pixel2Mesh by replacing VGG-16 with Resnet50 and integrating transformer blocks to better model global to local shape information for robust multi-scale 3D reconstruction. VGG-16 was replaced with Resnet50 to increase the stability of image feature extraction and to avoid the vanishing gradients problem through residual connections. The introduction of transformer encoder-decoder modules attend to varying levels of the GCN, enhancing long-range feature learning and contextual reasoning. Finally, we design new multi-scale loss functions tailored for transformers that leverage attention maps to improve vertex positioning and local surface geometry reconstruction.

We evaluate SculptFormer on a subset of 13 object categories taken from the larger ShapeNetCore dataset [1] which contains around 48,600 3D models across 55 object categories. These 13 object categories were previously used to evaluate Pixel2Mesh [10], enabling direct comparisons of our performance gains against their mesh deformation approach using standard metrics like Chamfer distance and F-Score. The multi-representation support with ground truth meshes, voxels, point clouds, and renderings also facilitates comprehensive geometric evaluations beyond overall shape similarity. Additionally, ShapeNetCore's scale and breadth test generalization across diverse 3D geometries, allowing for rigorous validation of our transformer architecture's multi-scale 3D understanding capabilities while situating our results in the context of prior work.

2.2. Hierarchical Transformer Modules

We propose a hierarchical design with multiple transformer modules operating at different scales to effectively integrate both global and local shape information. The first is the global transformer module, which applies a multi-headed self-attention mechanism across all mesh vertices and image features extracted after the third convolutional block in Resnet50. This self-attention allows each vertex to attend to representations from all other vertices in the mesh, aggregating global context to help maintain overall 3D structure, proportions, and vertex relationships. The outputs of this global self-attention are then passed through a graph residual block containing graph convolutional layers. This enhances the global features by also incorporating local information from each vertex's neighboring regions on the mesh surface.

After this initial coarse processing by the global transformer, an intermediate transformer module further bridges global and local contexts through a dual-level self-attention scheme. It captures not just single vertex relationships, but relationships between a vertex and clusters of neighboring vertices. This multi-scale attention allows the model to connect localized geometries to the broader shape structure. The intermediate transformer also incorporates positional encodings to maintain spatial consistency as the 3D geometry gets progressively refined.

Finally, the local transformer module operates at the most local level to recover intricate geometric details. It uses vector attention, a self-attention variant that efficiently scales to larger mesh resolutions by attending within local neighborhoods around each vertex instead of globally. This localized self-attention mechanism precisely adjusts vertex positions based on their surrounding context, incrementally adding details like sharp edges, corners, and thin structures missed by previous coarser stages. The hierarchical transformer architecture allows SculptFormer to coherently integrate multi-scale shape information, capturing the global 3D structure and intricate local geometry from just the 2D image input.

3. Experimental Results And Analysis

Implementation Details Our input images have dimensions of 127×127 pixels with no backgrounds. We use the Adam optimizer with a batch size of 8, an initial learning rate of $1e^{-4}$, a learning rate decay of 0.3 every 30 epochs. We train all modules, including Resnet50 and all transformers, end-to-end for 90 epochs. The resulting ellipsoid outputted by the last transformer block consists of 8192 vertices, rather than the 2466 vertices in the original Pixel2Mesh paper. The training process consumed around 16 hours on 8 A100 GPUs.

Dataset To evaluate the quality of our 3D mesh reconstructions, we report results on the aforementioned subset of 13 object categories taken from the ShapeNetCore dataset using widely-adopted quantitative metrics. We follow the standard dataset splits, using 70 percent for training, 10 percent for validation, and the remaining 20 percent for testing.

Metrics Our first key metric is Chamfer Distance (CD), which measures the relative distance of points sampled from the predicted 3D mesh surface to points on the surface of the ground truth object. It is calculated as the average of two symmetric distance terms - the sum of squared distances from each predicted mesh point to its nearest neighbor on the ground truth surface, and vice versa. A lower Chamfer Distance indicates the predicted mesh vertices are in close proximity to the true surface, capturing precise geometric details faithfully.

We also report the F-Score (F1), which evaluates the overall similarity between the predicted and ground truth 3D shape volumes. It is computed based on the intersection-over-union (IoU) of the predicted and ground truth occupancies, essentially measuring how well the predicted mesh aligns with the true solid shape as opposed to just the surface. Higher F1 scores denote better overall shape coherence and completeness in the 3D reconstruction. While Chamfer Distance focuses specifically on surface accuracy, and F-Score captures shape similarity more holistically, using both metrics in conjunction allows us to comprehensively analyze our method's ability to reconstruct high-fidelity 3D meshes preserving intricate geometric details as well as plausible global structures from just single-view 2D image input.

Qualitative Figure 2 showcases representative qualitative results directly comparing the 3D mesh reconstructions from our SculptFormer approach against the original Pixel2Mesh framework [1] and the ground truth shapes from ShapeNetCore. Across all examples spanning different object categories like airplanes, tables, cabinets and lamps, we can clearly see that SculptFormer generates significantly higher-fidelity 3D meshes better preserving intricate geometric details.

For the airplane model, our reconstruction faithfully recovers the thin wings, engine nacelles and horizontal stabilizers that appear smoothed over in Pixel2Mesh's output. The table example highlights SculptFormer's ability to capture precise surface patterns like the ribbed tabletop design. Our method also excels at reconstructing objects with complex curved geometries like the cabinet model, producing much crisper edges and handles compared to Pixel2Mesh. Beyond improved detail preservation, our 3D mesh predictions also exhibit more plausible and coherent global structures adhering to the overall shape proportions. This can be seen in the lamp example, where Pixel2Mesh's output appears distorted, while SculptFormer faithfully reconstructs the accurate curved geometry of the lamp base and shade components. Crucially, our transformer-based coarse-to-fine architecture allows seamlessly integrating fine detail recovery within a coherent global 3D understanding, overcoming trade-offs in previous techniques. The model is able to dynamically focus on different shape scales - first approximating an overall plausible 3D structure grounded in

Figure 2. Qualitative results of meshes reconstructed using Sculpt-Former

Figure 3. Qualitative results of meshes from Pixel2Mesh. Left two columns are ground truth while right two columns are outputs

	SculptFormer (ours)	Pixel2Mesh
Vessel	0.228	0.670
Cabinet	0.169	0.381
Table	0.172	0.498
Chair	0.170	0.610
Rifle	0.274	0.453
Plane	0.139	0.477
Speaker	0.158	0.739
Lamp	0.198	1.295
Phone	0.217	0.421
Sofa	0.155	0.490
Bench	0.218	0.624
Display	0.253	0.755
Car	0.125	0.268

Table 1. Comparison of Chamfer Distance (lower is better)

the image context, before progressively adding precise local geometric details guided by both the image features and its growing 3D shape understanding.

Quantitative The quantitative results highlight Sculpt-Former's significant geometric accuracy gains over the Pixel2Mesh baseline across most object categories in the challenging ShapeNetCore dataset. Looking at the Chamfer Distance (CD) results in Table 1, which directly measure mesh surface precision, our method achieves substantially lower CD values indicating much higher-fidelity detail preservation. For geometric structures like airplane wings (CD 0.139 vs 0.477) and thin components like rifle barrels (0.274 vs 0.453), SculptFormer demonstrates over 60 percent lower CD compared to Pixel2Mesh.

For several categories like rifles (0.4664 vs 0.8347) and phones (0.5505 vs 0.8286), we also see SculptFormer outperforming Pixel2Mesh in terms of the F-Score metric. However, it's important to note that the F-Score calculation was conducted on a limited sample of just 5 examples per object category due to time and computational constraints during our evaluations. This very small sample size may not adequately capture the full performance distribution across the dataset. With such a small sample, even just one or two failure cases with poor overlap could significantly skew the averaged F-Score downwards for that category. This sampling issue is especially pronounced for categories with higher intra-class shape variation like rifles and phones which can exhibit diverse geometries. In contrast, our Chamfer Distance results demonstrate clear advantages for SculptFormer in accurately reconstructing precise surface geometry details for these same categories. Chamfer Distance directly measures averaged vertex-to-surface distances, making it less sensitive to sampling issues compared to the volume intersection metric used for F-Scores. It's likely that with a larger, more representative sample, the F-Scores for these categories would better align with the geometry precision indicated by our Chamfer Distance numbers. Unfortunately, we were limited by computational resources in calculating scores over more examples per category for our evaluation. Promisingly, for smoother, more chunk-like object categories where we expect less variation across samples, like airplanes (0.7915 vs 0.8238) and cars (0.7801 vs 0.8415), our F-Scores are very competitive with Pixel2Mesh despite the sampling limits. This suggests the sampling issue is less pronounced when intra-class geometries are more consistent.

4. Conclusion

We present a transformer-boosted 3D mesh reconstruction framework that builds upon the Pixel2Mesh method by adding hierarchical transformer blocks to effectively combine localized geodesic information from each ver-

CVPR #*****

Category	SculptFormer (ours)	Pixel2Mesh
Vessel	0.5500	0.6999
Cabinet	0.6863	0.7719
Table	0.6505	0.7920
Chair	0.6808	0.7042
Rifle	0.4664	0.8347
Airplane	0.7915	0.8238
Speaker	0.7161	0.6561
Lamp	0.6780	0.6150
Phone	0.5505	0.8286
Sofa	0.7162	0.6983
Bench	0.6187	0.7186
Display	0.5749	0.6701
Car	0.7801	0.8415

Table 2. Comparison of F-score (higher is better)

tex's neighboring regions with global context. Our results show an improved performance as compared to the original Pixel2Mesh. At the time of writing, two similar architectures, T-Pixel2Mesh [8] and InstantMesh [6], which also utilize transformers and a novel Large Reconstruction Model (LRM) based architecture to improve mesh generation quality have also very recently released. We hope our work encourages future work that utilizes other transformer-based architectures for improved 3D reconstruction models.

5. Individual Contributions

Evan ran experiments to replicate Pixel2Mesh's Chamfer distance results for each of the chosen 13 object categories, while Shrika ran experiments to replicate Pixel2Mesh's F-score results for each of the chosen 13 object categories. We both worked together to make significant changes to the original Pixel2Mesh model architecture and incorporate global, intermediate, and local transformer blocks as outlined in our paper. Shrika focused on changing the network for feature extraction from VGG-16 to Resnet50 and attaching the global transformer to the Resnet50 and the underlying graph convolutional network (GCN). Evan then focused on attaching the intermediate and local transformer modules to work with the global module, Resnet, and the GCN. We tested each of our respective portions of work, ensuring that each incremental addition would work with the rest of the architecture. Once our architecture was set up correctly, we each ran multiple experiments each day with varying configurations of batch size, learning rate, learning rate decay, and various other parameters when deemed necessary. We each had to run many experiments initially since our runs would fail prematurely. Later on, we could only train a few times a day since the training runs would take several hours, and we would terminate them prematurely if results did not appear to be promising. Shrika prepared code to visualize the qualitative results. Evan worked on scripting portions of the testing process and setting up the data. We both tried to collect qualitative metrics from the Pixel2Mesh implementation, but due to a bug when using their visualizer that we could not figure out during result generation, we could not provide those results, even though quantitative results were replicated. Thus, qualitative results were taken directly from the Pixel2Mesh paper. All sections of the paper were co-written, revised, and looked over by both of us. We each created one table of results. We both worked on creating the figure for our approach.

References

[1] Leonidas Guibas et.al Angel X. Chang, Thomas Funkhouser. Shapenet: An information-rich 3d model repository. In *Eurographics Workshop on 3D Object Retrieval*, 2015. 2, 3

[2] Zhenwei Bian Jun Li-Kai Xu Chengjie Niu, Yang Yu. Weakly supervised part-wise 3d shape reconstruction from single-view rgb images. In *Computer Graphics Forum*, 2020. 2

[3] Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2

[4] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[5] Leonidas Guibas Haoqiang Fan, Hao Su. A point set generation network for 3d object reconstruction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[6] Yiming Gao Xintao Wang-Shenghua Gao Ying Shan Jiale Xu, Weihao Cheng. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 5

[7] Florian Golemo Jérôme Parent-Lévesque David Vazquez Derek Nowrouzezahrai Aaron Courville Sai Rajeswar, Fahim Mannan. Pix2shape: Towards unsupervised learning of 3d scenes from images using a view-based representation. In *International Journal of Computer Vision*, 2020. 2

[8] Keke He Junwei Zhu-Ying Tai Chengjie Wang Yinda Zhang Yanwei Fu Shijie Zhang, Boyan Jiang. T-pixel2mesh: Combining global and local transformer for 3d mesh generation from a single image. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. 5

[9] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[10] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2