SculptFormer

Transformer Boosted 3D Mesh Reconstruction from a 2D Image

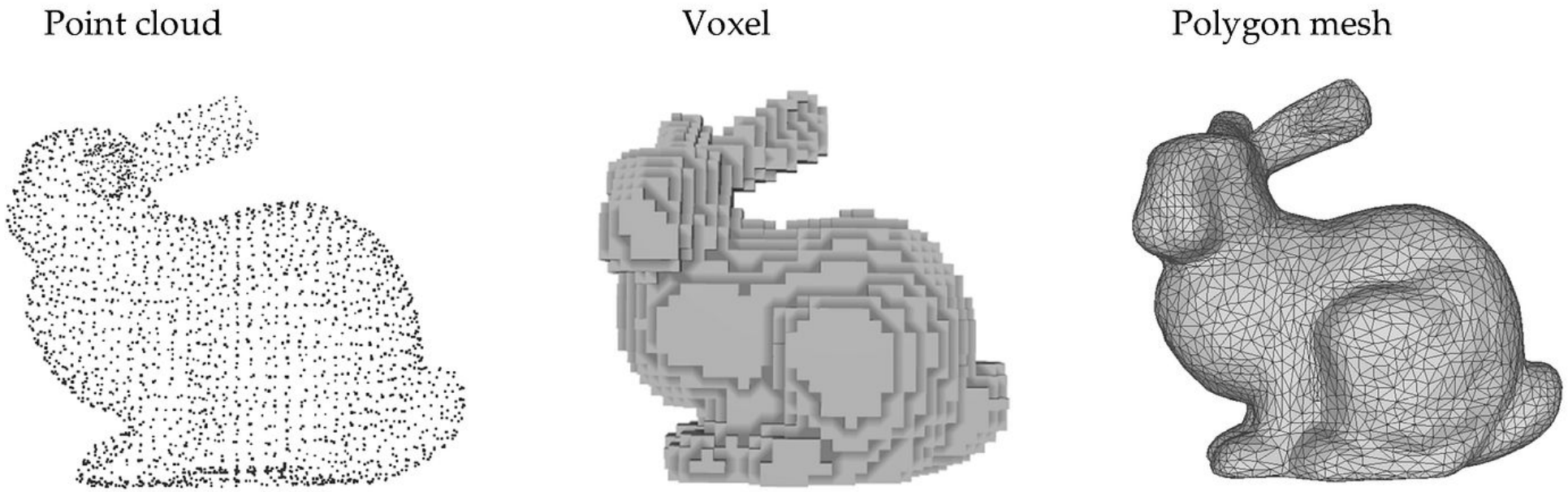
Evan Kim & Shrika Eddula

May 2024

Massachusetts Institute of Technology

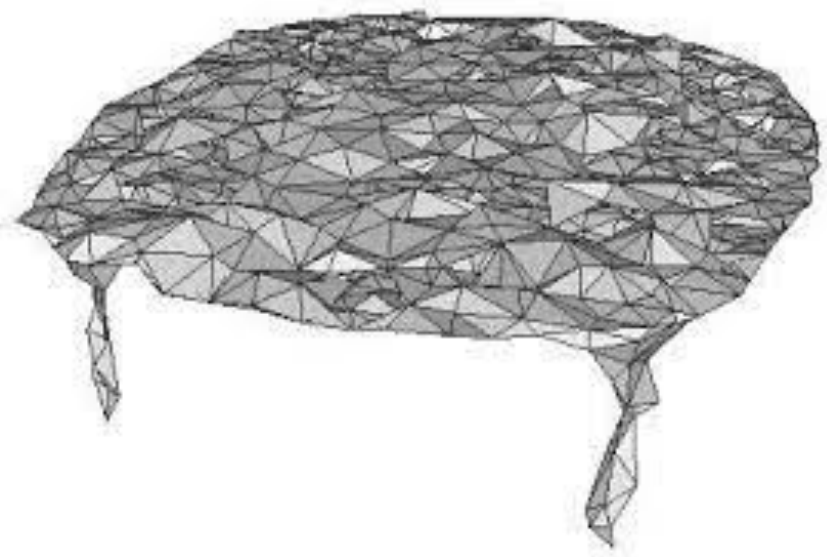
Background

- ➔ 3D reconstruction from single 2D image → object manipulation tasks for robotic systems, AR/VR experiences, etc
- ➔ Point Cloud, Voxel, and Mesh based methods

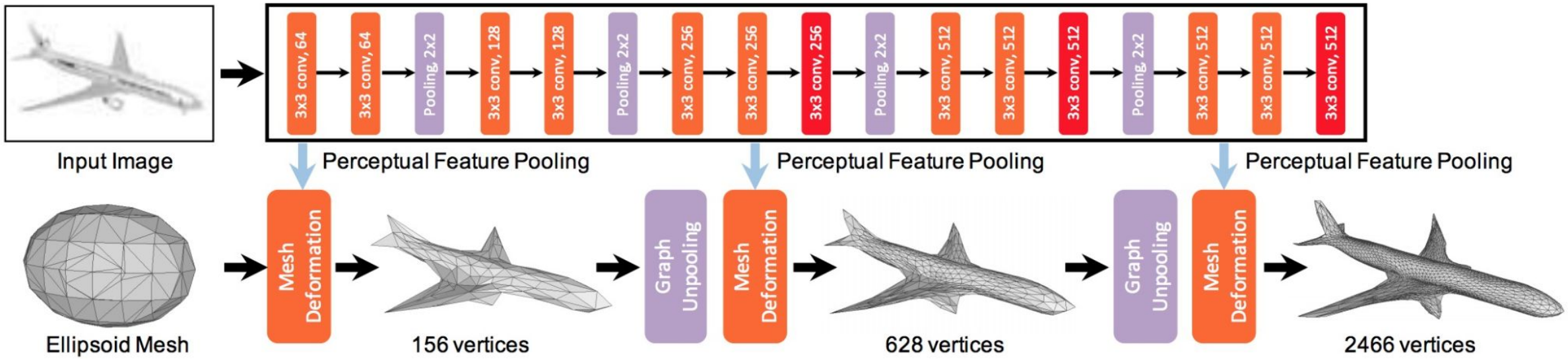


Limitations in Current Methods

- ➔ **Challenge:** simultaneously capturing accurate holistic shape information AND intricate local geometric details
- ➔ 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
- ➔ Recently: **transformers** gaining attention



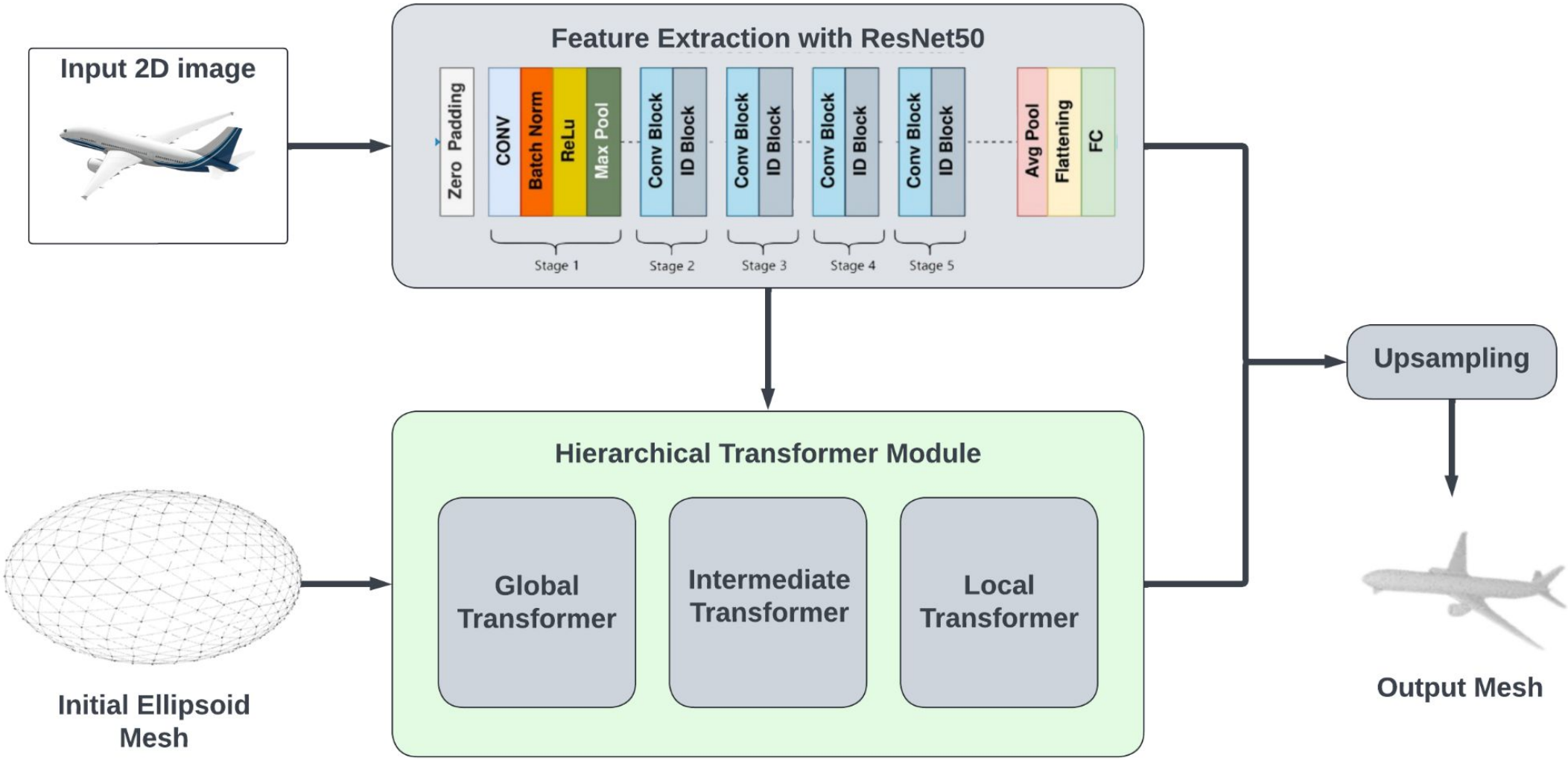
Proposed Method



Pixel2Mesh: graph convolutional neural network (GCN) that deforms an initial ellipsoid mesh towards the target 3D shape, coarse-to-fine

Hierarchical Transformer Modules: boost existing Pixel2Mesh architecture with addition of transformer blocks

Proposed Method



Analysis

- **ShapeNet Core** data set
  48,600 3D models across 55 object categories
- Established benchmark, utilize a subset of objects taken from 13 of the 55 categories
- Chamfer Distance and F Score
- Direct comparisons of our performance against Pixel2Mesh

Qualitative Results

Pixel2Mesh

SculptFormer (ours)

Quantitative Results

	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 reconstruction accuracy using Chamfer Distance (lower is better)

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 reconstruction accuracy using F-score (higher is better)

Conclusions

- ➔ Transformer boosted 3D mesh reconstruction framework that builds upon the Pixel2Mesh method by adding a **hierarchical transformer module**

- ➔ **Improved** performance, especially in regards to **fine detail**

- ➔ Further encourage the path towards the **integration** of 3D reconstruction models with **transformer based** architecture