

MESC-3D: Mining Effective Semantic Cues for 3D Reconstruction from a Single Image

Supplementary Material

To provide a more comprehensive explanation of our method, this supplementary material includes detailed information on various aspects of our method:

- Training and Dataset Details
- Complexity Analysis
- More Visualization Results
- Algorithm MESC-3D Explanation

1. More Implementation Details

Training for Two Stages. For the 3D input, we follow the experimental settings of 3DAttriFlow, uniformly sampling $N_p = 2048$. Our learnable text prompts and 3D reconstruction are trained in two separate stages, with both stages using only data from the training and validation sets, excluding any data from the test set. It is worth noting that our approach adopts a all-categories strategy., unlike methods such as PC² and BDM, which rely on single-category training for diffusion.

Generalization Capability Experiments. For the generalization capability experiment, we replace the image encoder ResNet18 with a CLIP large model and DPT, fine-tuned on ShapeNet but tested on Pix3D. Finally, we conduct robustness testing on Pix3D, demonstrating that our network effectively mining semantic information for 3D shape reconstruction.

Zero-shot Capability Experiments. The learnable text prompt is trained exclusively on the ShapeNet dataset (comprising 13 categories), and subsequently embedded to provide prior guidance for the reconstruction of previously unseen categories.

2. Dataset Details

We continue testing qualitative results, parameter numbers, and inference time on a subset of ShapeNet. In robustness experiments, we not only test on Pix3D but also download some online photos for 3D reconstruction, further validating the robustness and efficiency of our network.

3. Complexity Analysis

As shown in Tab. A, the comparison results indicate that the inference speed of diffusion models is significantly slower than ours, and they also use the most parameters. Compared to 3DAttriFlow, MESC3D performs on par with prior work. Although incorporating text prompt encoding naturally slows down inference slightly, our CDL2 metric greatly exceeds theirs. We also conducted an impact test

on the number of point clouds as seen in Tab. B. When increasing the number of point clouds from 2048 to 8192, the effect on our training and inference times was minimal.

Table A. Complexity and inference time of different methods. w/o and w represent without and with text prompt respectively.

Methods	Params	Infer time	Avg- $CD\ell_2 (\times 10^3)$
Point-e	80.94M	55.215s	155
3DAttriFlow	20.92M	0.117s	4.08
PC ²	27.65M	2.800s	5.39
BDM-B	49.71M	7.602s	5.3
Ours (w/o)	24.05M	0.165s	3.69
Ours (w)	24.97M	0.548s	3.22

Table B. Impact of the number of point cloud on inference time.

Number of points	Infer time
Ours(w/o)2048	0.165s
Ours(w/o)8192	0.309s

4. More Visualization Results

We offer additional visualization results on the ShapeNet dataset that demonstrate the superior performance of our method in recovering occluded regions from a single image. For example, our method successfully reconstructs the fully occluded sofa cushion as seen in Fig. C, and the recovery of the truck bed is remarkable. Additionally, we excel in categories with objects that have fine details, such as the tail of the airplane and the shape recovery of the fighter jet as seen in Fig. B. Compared to the diffusion-based method, our network has three main advantages:

- Accurate foreground-background identification, ensuring the correct object is reconstructed from a single image with a higher reconstruction category accuracy.
- Effective utilization of semantic information to guide the 3D reconstruction.
- Consistency in results. Repeated inputs of the same image yield consistent output, while Point-E produces varied results each time.

Fig. A illustrates the zero-shot capability introduced by learnable text prompt.

The detailed steps and implementation of the MESC-3D algorithm are provided in Algorithm 1. In summary, our model demonstrates robust performance.

Algorithm 1 MESC-3D: Mining Effective Semantic Cues for 3D Reconstruction from a Single Image**Input:** I (image), P (point cloud)**Output:** P_{pred}

```

1: Extract image features:  $\mathbf{I}_{\text{feat}} = \text{ResNet18}(I)$ 
2: Extract point cloud features:  $\mathbf{P}_{\text{feat}} = \text{PointMAE}(P)$ 
3: Initialize  $Q_0$  as random query values
4: for each layer  $t = 1$  to  $T$  do
5:   if  $t$  is even then
6:     Set  $Q^t = \mathbf{P}_{\text{feat}}, K^t = V^t = \mathbf{I}_{\text{feat}}$ 
7:   else
8:     Set  $Q^t = \mathbf{I}_{\text{feat}}, K^t = V^t = \mathbf{P}_{\text{feat}}$ 
9:   end if
10:  Perform attention:  $\mathbf{F}_{\text{fusion}}^t = \text{Attention}(Q^t, K^t, V^t)$ 
11:  Update query:  $Q^{t+1} = \mathbf{F}_{\text{fusion}}^t$ 
12: end for
13: Initialize  $\text{dec\_dim} = [768, 512, 256, 128, 64, 32]$ 
14: for each layer  $l = 1$  to  $L$  do
15:  Compute downsampled features:
     $\mathbf{F}_{\text{down}}^l = \text{conv}_l(\mathbf{F}_{\text{fusion}}^{l-1})$ 
16:  Select features:
     $\mathbf{F}_{\text{select}}^l = \text{map}_l(\mathbf{F}_{\text{fusion}}^{l-1})$ 
17:  Normalize and fuse features:
     $\mathbf{F}_{\text{next}}^l = \text{AdaptivePointNorm}(\mathbf{F}_{\text{down}}^l, \mathbf{F}_{\text{select}}^l)$ 
18:  Update:  $\mathbf{F}_{\text{fusion}}^l = \mathbf{F}_{\text{next}}^l$ 
19: end for
20:  $\mathbf{F}_{\text{final}} = \mathbf{F}_{\text{next}}^L$  {Final fused features}
21: MLP for Point Cloud Reconstruction:
22:  $\mathbf{P}_{\text{pred}} = \text{MLP}(\mathbf{F}_{\text{final}})$  {Apply MLP to map to point cloud features}
23: return  $\mathbf{P}_{\text{pred}}$ 

```

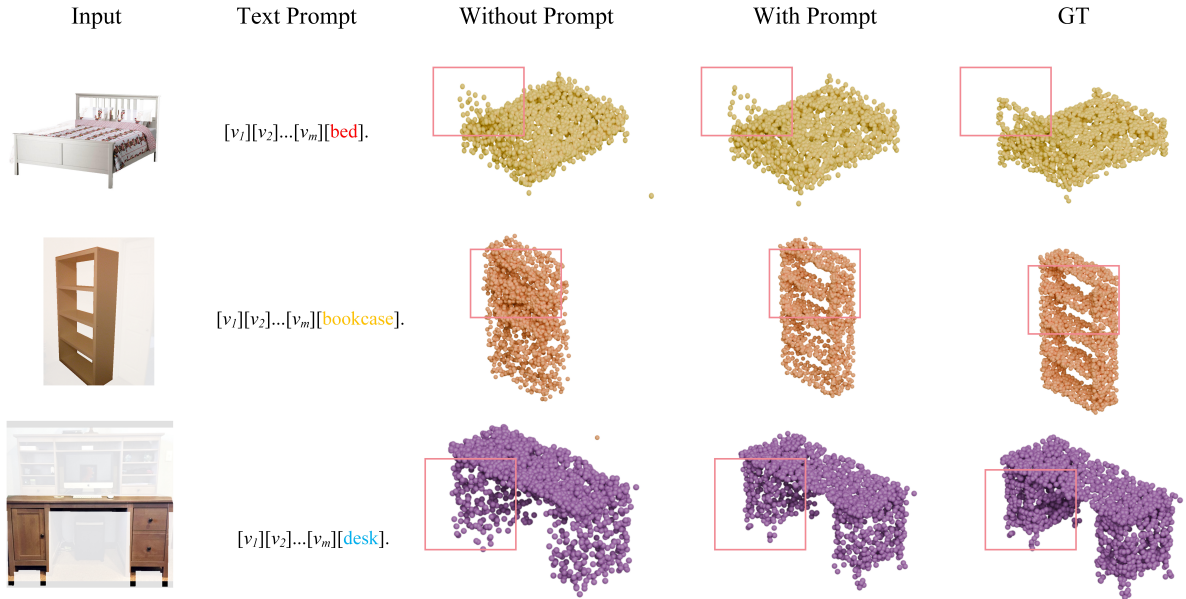


Figure A. Demonstration of the zero-shot ability of learnable text prompt, enabling detailed 3D shape reconstruction for unseen object categories.

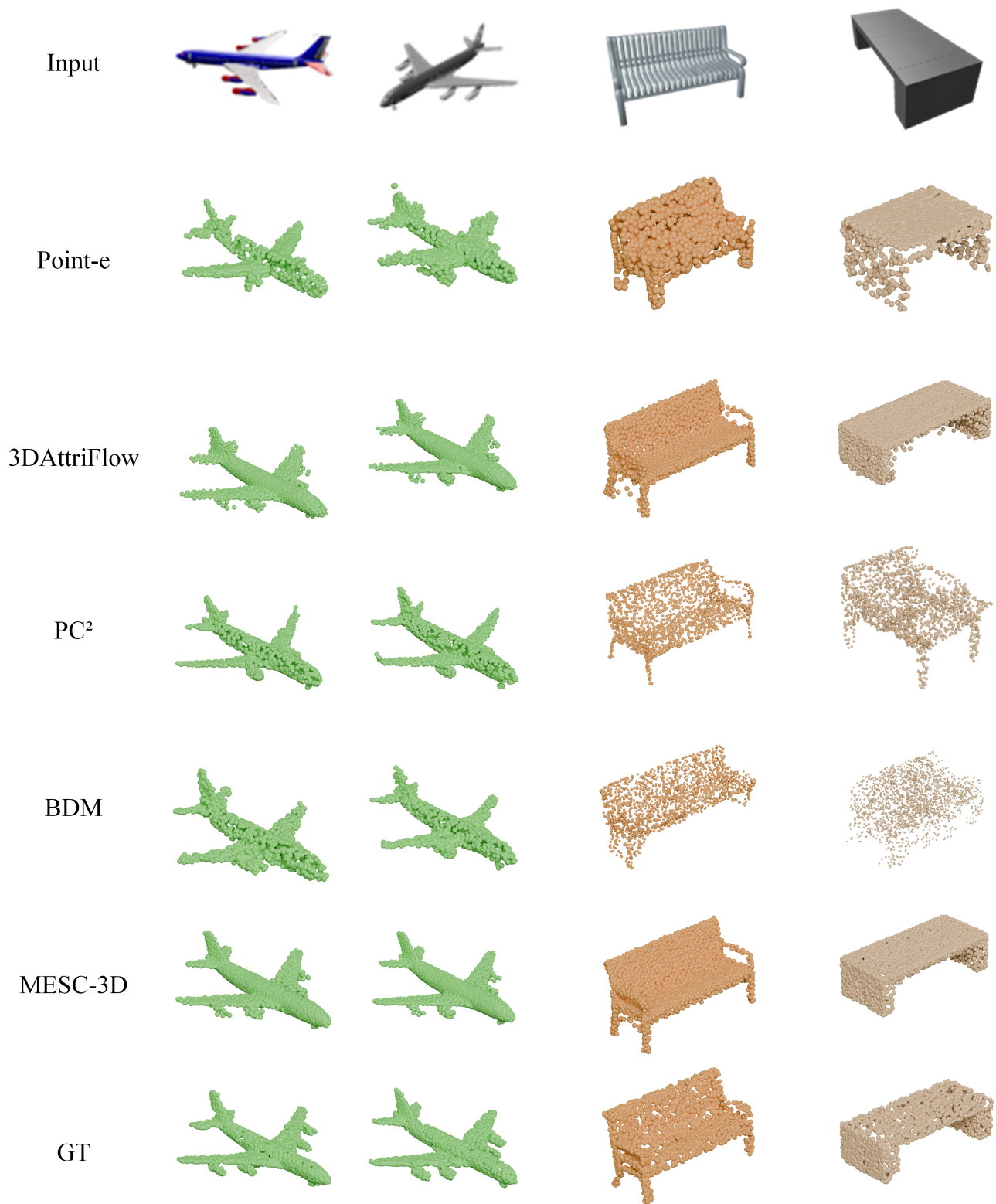


Figure B. Visual comparison of 2D-to-3D reconstruction results with different methods on “airplane” and “bench” in ShapeNet dataset.

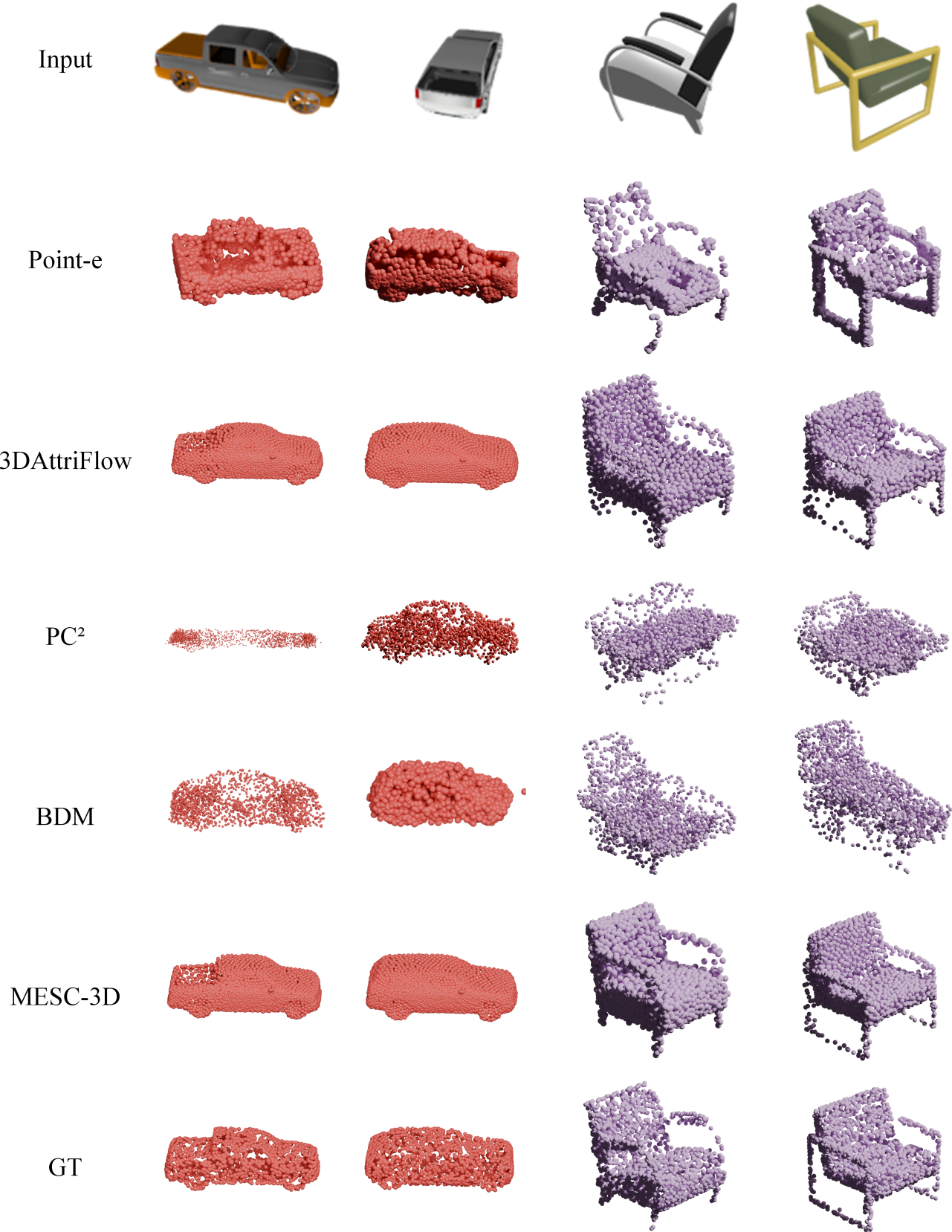


Figure C. Visual comparison of 2D-to-3D reconstruction results with different methods on “car” and “chair” in ShapeNet dataset.

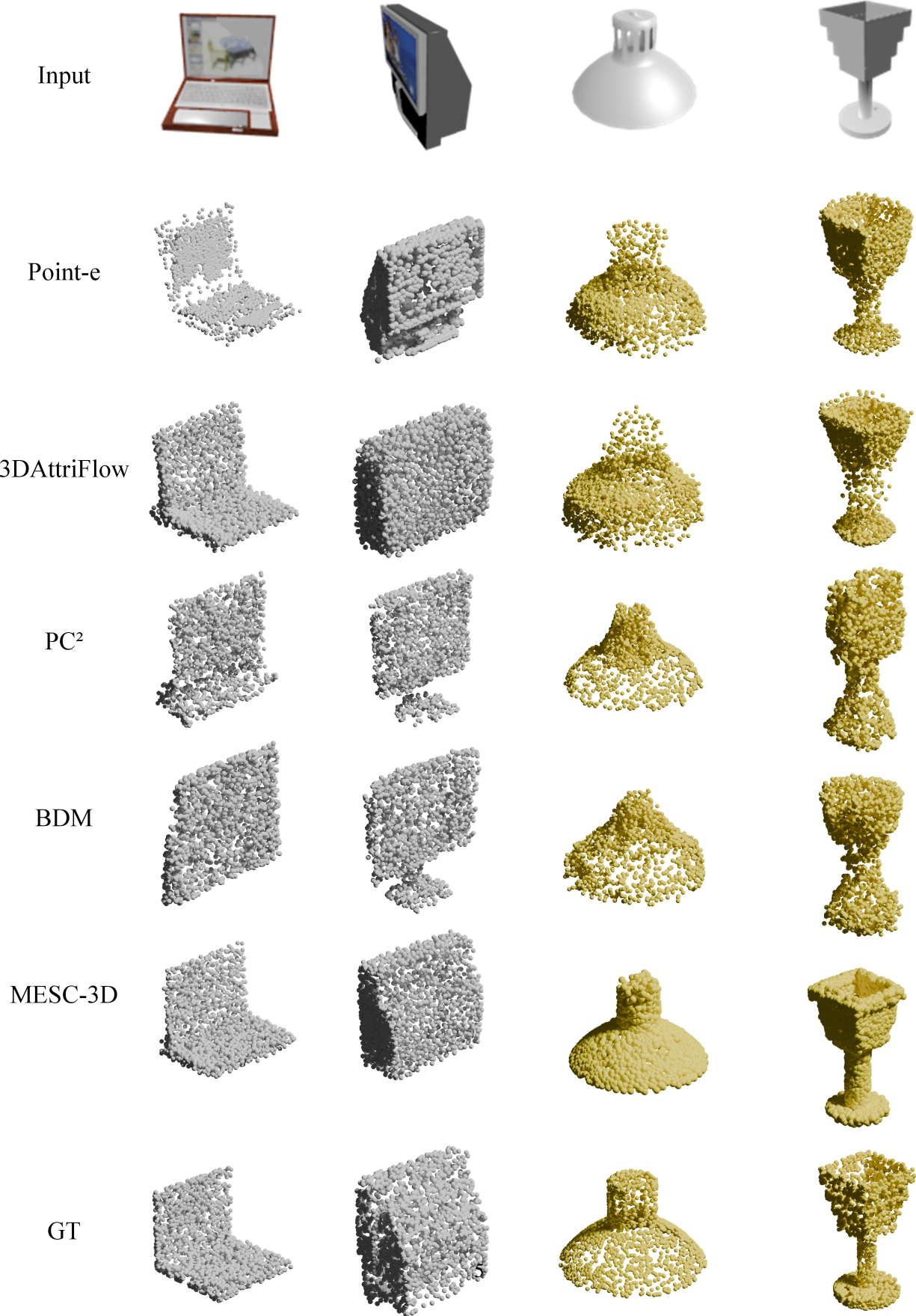


Figure D. Visual comparison of 2D-to-3D reconstruction results with different methods on “display” and “lamp” in ShapeNet

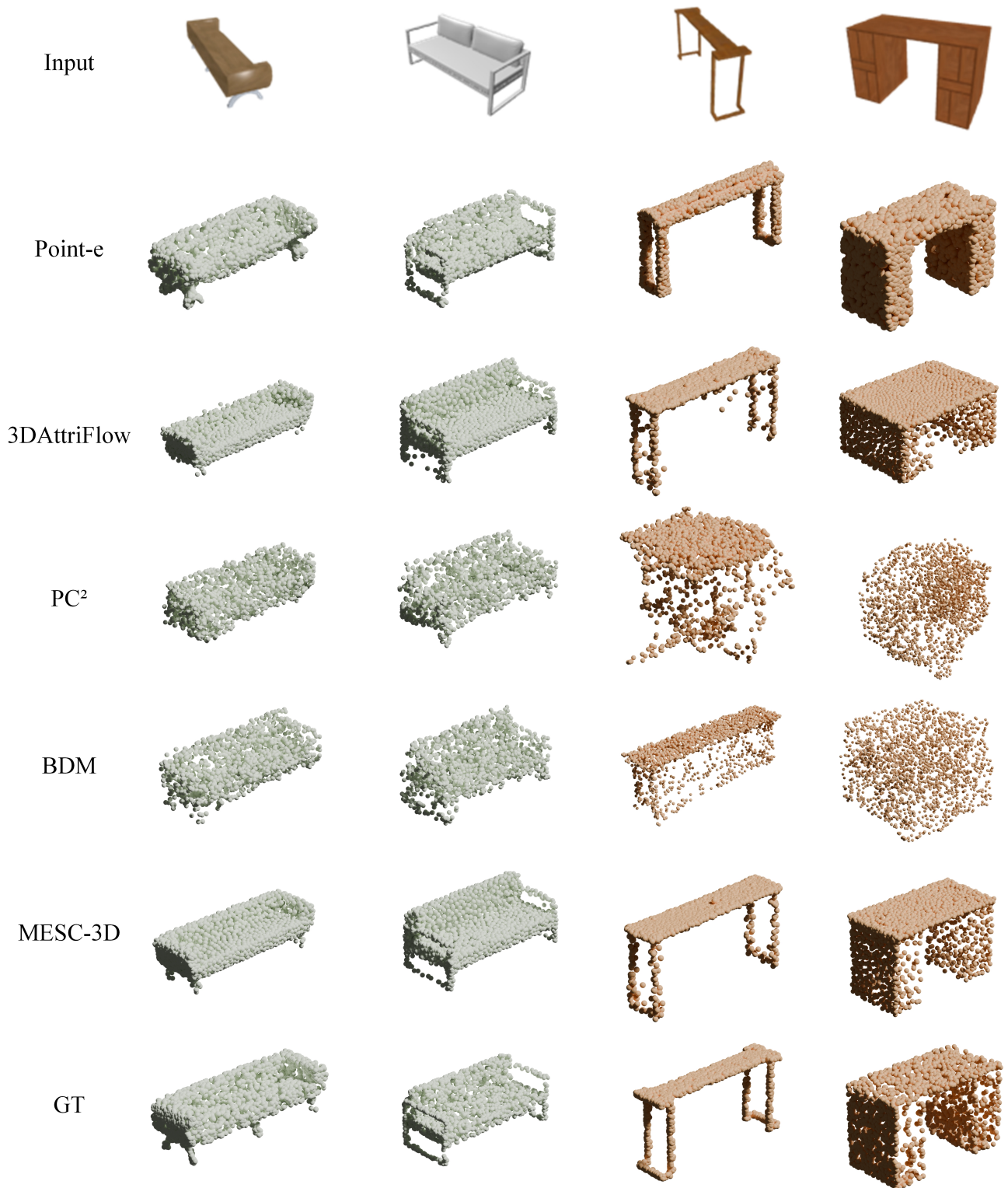


Figure E. Visual comparison of 2D-to-3D reconstruction results with different methods on “sofa” and “table” in ShapeNet dataset.

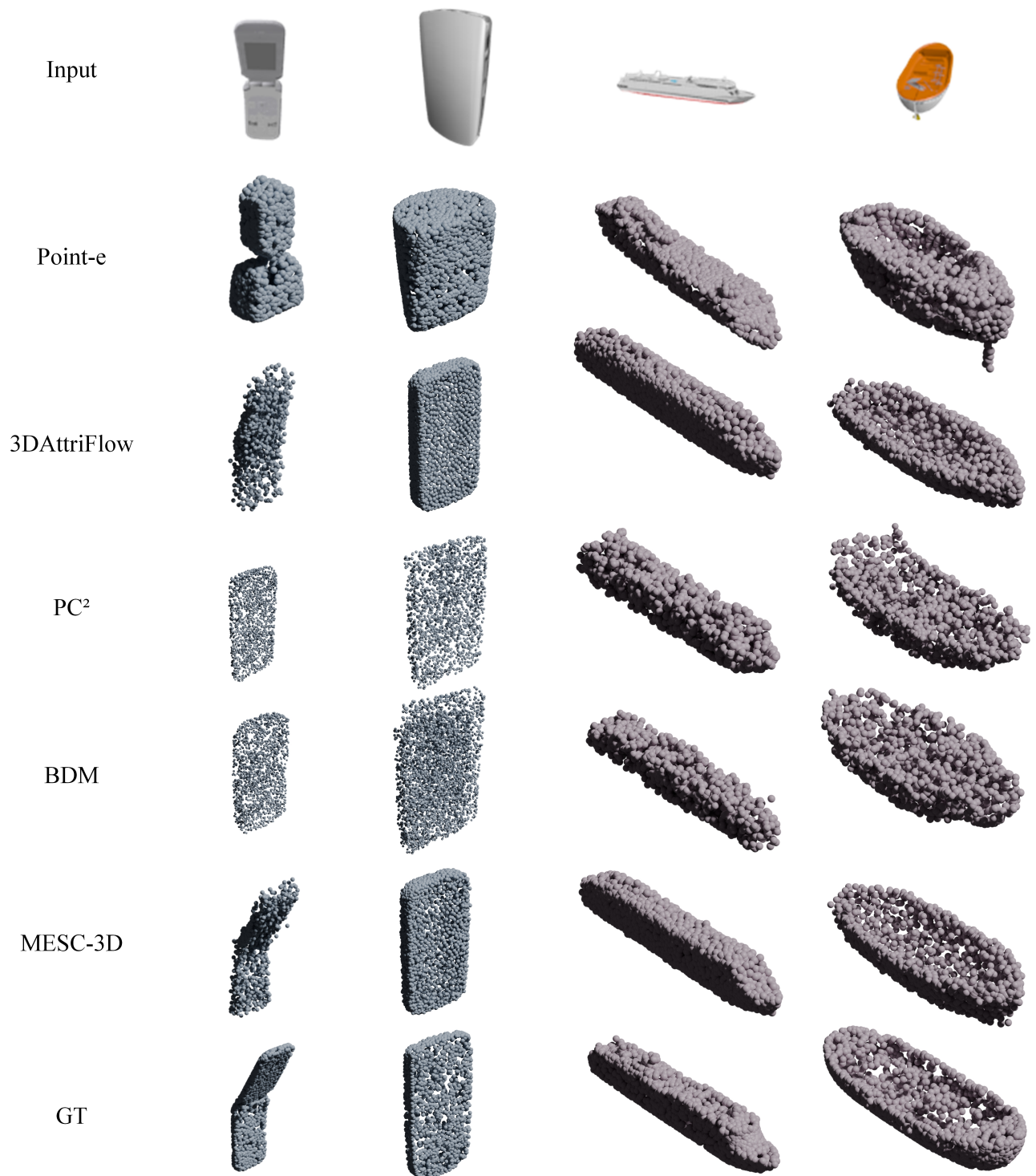


Figure F. Visual comparison of 2D-to-3D reconstruction results with different methods on “telephone” and “vessel” in ShapeNet dataset.