Supplementary Material for CVPR 2025 Paper: Learning Partonomic 3D Reconstruction from Image Collections

Xiaoqian Ruan[1], Pei Yu[2], Dian Jia[1], Hyeonjeong Park[1], Peixi Xiong[3], Wei Tang[1]
[1]University of Illinois Chicago, [2]Microsoft, [3]Intel
{xruan9, djia7, hpark233, tangw}@uic.edu, pei.yu@microsoft.com, peixi.xiong@intel.com

1. More Details on Loss Functions

Following the previous works [7, 10], our regularization loss $\mathcal{L}^{\text{reg}}$ consists of five terms:

$$\lambda^{\text{norm}}\mathcal{L}^{\text{norm}} + \lambda^{\text{lap}}\mathcal{L}^{\text{lap}} + \lambda^{\text{perc}}\mathcal{L}^{\text{perc}} + \lambda^{\text{nbr}}\mathcal{L}^{\text{nbr}} + \lambda^{\text{pose}}\mathcal{L}^{\text{pose}} \tag{1}$$

where the balancing hyper-prameters λ^{norm}, λ^{lap}, λ^{perc}, λ^{nbr}, and λ^{pose} are set to 0.01, 0.01, 10, 1, and 0.2, respectively.

The normal consistency loss [3] $\mathcal{L}^{\text{norm}}$ is defined as:

$$\mathcal{L}^{\text{norm}} = \sum_{(i,j)\in\mathcal{E}} (1 - \frac{\boldsymbol{n}_i \cdot \boldsymbol{n}_j}{||\boldsymbol{n}_i||_2||\boldsymbol{n}_j||_2}) \tag{2}$$

where $\mathcal{E}$ represents the set of edges shared by two neighboring faces, and $\boldsymbol{n}$ is a face's normal vector.

The Laplacian smoothing loss [11] is defined as:

$$\mathcal{L}^{\text{lap}} = \sum_i \left|\left| \boldsymbol{v}_i - \frac{1}{|\mathcal{N}(i)|} \sum_{j\in\mathcal{N}(i)} \boldsymbol{v}_j \right|\right|_2^2 \tag{3}$$

where $\boldsymbol{v}$ and $\mathcal{N}(i)$ represent a vertex and a set of neighboring vertices, respectively.

The perceptual loss [15] $\mathcal{L}^{\text{perc}}$ is defined as:

$$\mathcal{L}^{\text{perc}} = \left\|\phi(\boldsymbol{I}') - \phi(\boldsymbol{I})\right\|_2^2 \tag{4}$$

where $\phi(\cdot)$ represents the output of the relu3_3 layer of a pretrained VGG16 [13].

Following the previous works [7, 10], we use neighbor reconstruction loss by exchanging the shape and texture attributes to enforce the consistency across instances within the same category.

$$\mathcal{L}^{\text{nbr}} = ||\boldsymbol{I}_i - \boldsymbol{I}'^{\text{tx}}_{j\to i}||_2^2 + ||\boldsymbol{I}_i - \boldsymbol{I}'^{\text{sh}}_{k\to i}||_2^2 \tag{5}$$

where samples j and k are the texture and shape neighbors [10] of sample i, respectively, and $\boldsymbol{I}'^{\text{tx}}_{j\to i}$ and $\boldsymbol{I}'^{\text{sh}}_{k\to i}$ are the reconstructed images by swapping the texture or shape code.

Category	Airplane	Car	Chair	Lamp	Table
ShapeNet	2831	5247	4744	1622	5956
ShapeNetPart	1908	654	2655	1004	2532

Table 1. Numbers of training instances in ShapeNet and ShapeNetPart. The training instances in ShapeNetPart are much fewer than those in ShapeNet, especially for the car category.

Finally, following the widely used camera multiplex [5, 10], $\mathcal{L}^{\text{pose}}$ is defined as:

$$\mathcal{L}^{\text{pose}} = \sum_k |\bar{p}_k - \frac{1}{K}| \tag{6}$$

where $\bar{p}_k$ is the frequency of the k-th pose candidate in a training batch. This loss encourages a uniform distribution on the pose estimates.

2. Data Generation

In this section, we briefly introduce the process of data generation for three datasets, including ShapeNetPart in Sec. 2.1, PartNet in Sec. 2.2, and CUB-200-2011 in Sec. 2.3. We will release all generated data along with our code.

2.1. ShapeNetPart

ShapeNetPart [4] contains significantly fewer training instances compared to ShapeNet [2], as shown in Table 1. For instance, there are only 654 training samples for the car category in ShapeNetPart, which is just one-tenth of the corresponding number in ShapeNet.

The data in ShapeNetPart is represented as point clouds, including point locations and their corresponding part labels. To generate the required data for training and evaluation, we first identify the corresponding mesh and point cloud in ShapeNet for each object in ShapeNetPart. To improve the quality of the generated part masks, we upsample the mesh using subdivision [8]. Next, we align the point cloud from ShapeNetPart with the mesh and point cloud

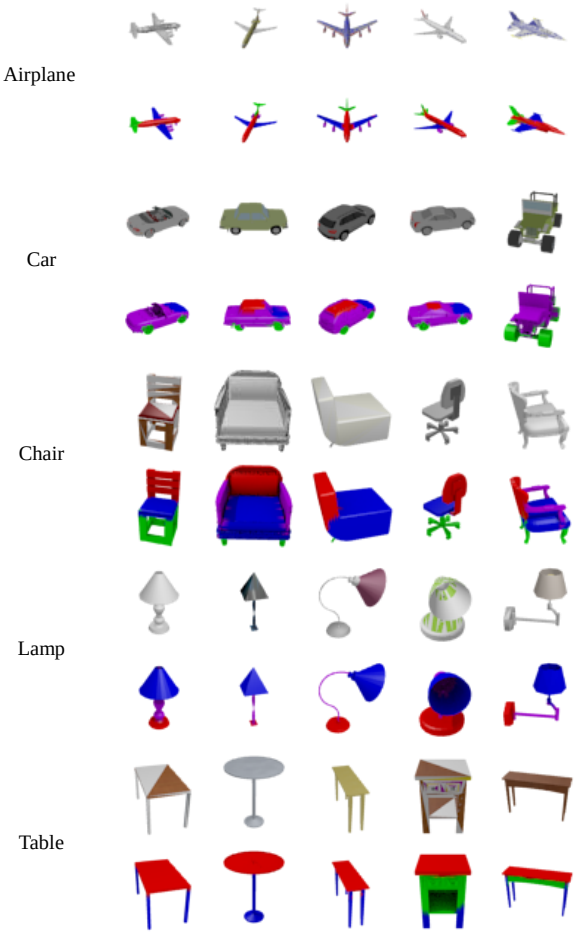

Figure 1. Examples of the input images and their corresponding part-segmented meshes across five categories in ShapeNetPart. Different colors represent different parts.

Category	Bottle	Bowl	Display	Knife	Mug
PartNet	315	131	633	221	138

Table 2. Numbers of training instances across five categories in PartNet.

from ShapeNet and transfer the part labels based on the distance between points. To generate training data, we render the image and object mask using the original mesh and render the part mask using the part-segmented mesh. The rendered images and the corresponding part-segmented meshes for example objects are illustrated in Figure 1. For evaluation purposes, we utilize the point cloud in ShapeNet with the transferred part labels, as it contains significantly more points than the point cloud in ShapeNetPart.

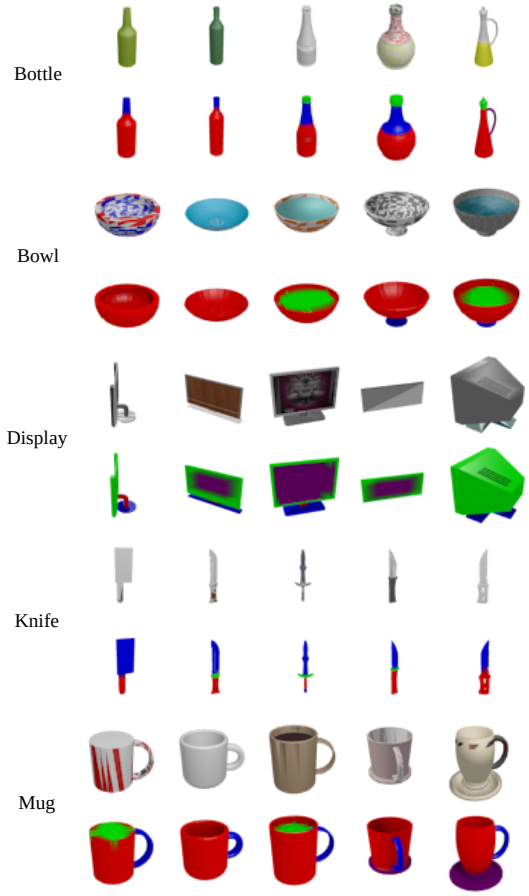

Figure 2. Examples of the input images along with the part-segmented meshes across the five categories in PartNet.

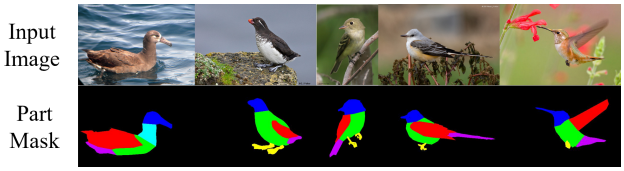

Figure 3. 2D part masks of CUB-200-2011. Different colors are used to represent different parts: head (blue), body (green), neck (cyan), wing (red), tail (purple), and legs (yellow).

2.2. PartNet

Table 2 shows the numbers of training instances for five PartNet [9] categories used in our experiments. PartNet contains significantly fewer samples than ShapeNetPart.

Since PartNet already provides the high-quality ground truth, including point locations and part labels, we only need to generate the 2D images, object masks, and part

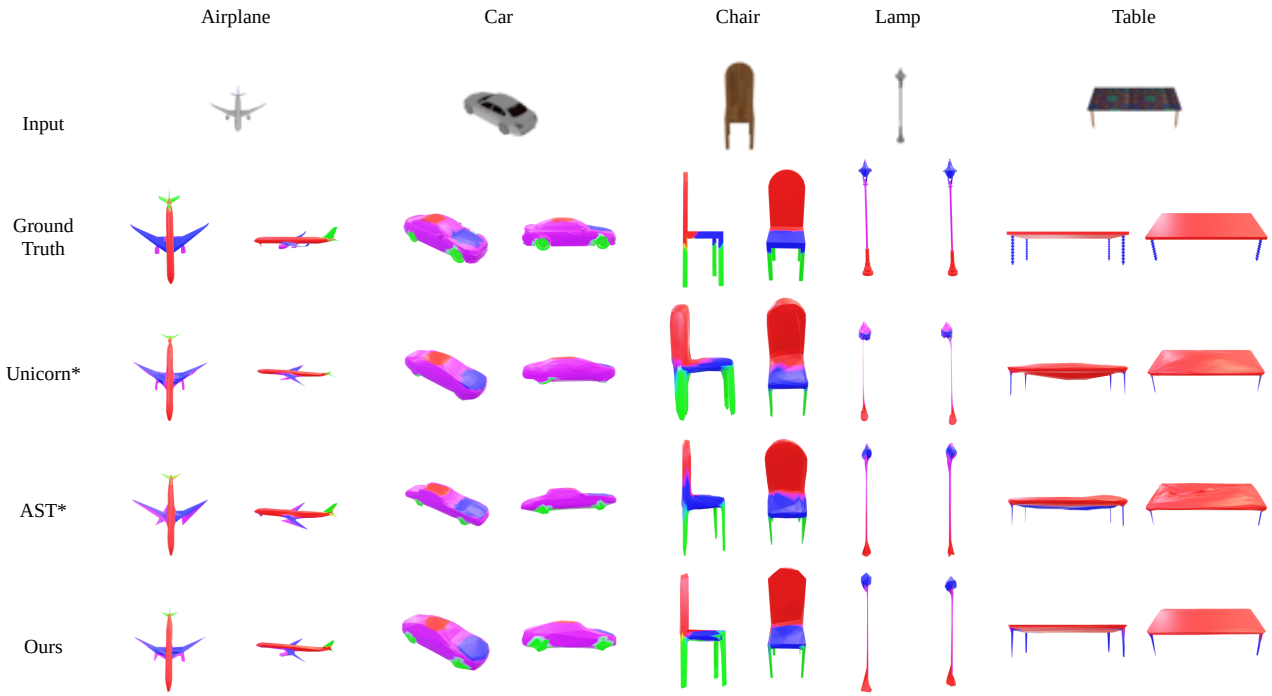

Figure 4. Qualitative results with five categories in ShapeNetPart based on three approaches. Unicorn [10] and AST [7] are state-of-the-arts methods for learning whole object reconstruction from image collections. We extend them for partonomic reconstruction.

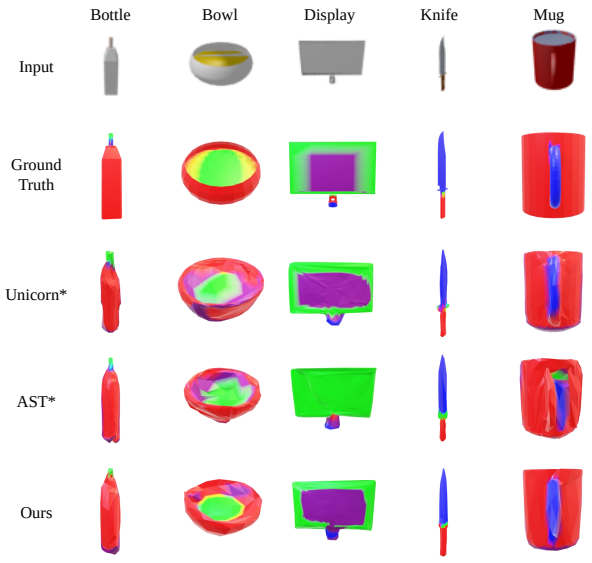

Figure 5. Qualitative results across five categories on PartNet with three methods. Unicorn [10] and AST [7] are state-of-the-art methods for learning whole object reconstruction from image collections. We extend them for partonomic reconstruction.

masks required for training. Similar to ShapeNetPart, we

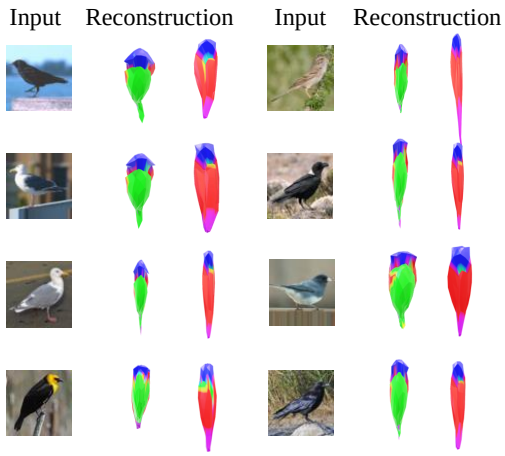

Figure 6. Additional qualitative results showing front and back views on CUB-200-2011. The input images of these examples are the same as those in Figure 5 in the main paper.

first identify the corresponding mesh in ShapeNet for each object in PartNet and upsample the mesh using subdivision [8]. Then, we align the point cloud from PartNet with the mesh in ShapeNet and transfer part labels to the correspond-

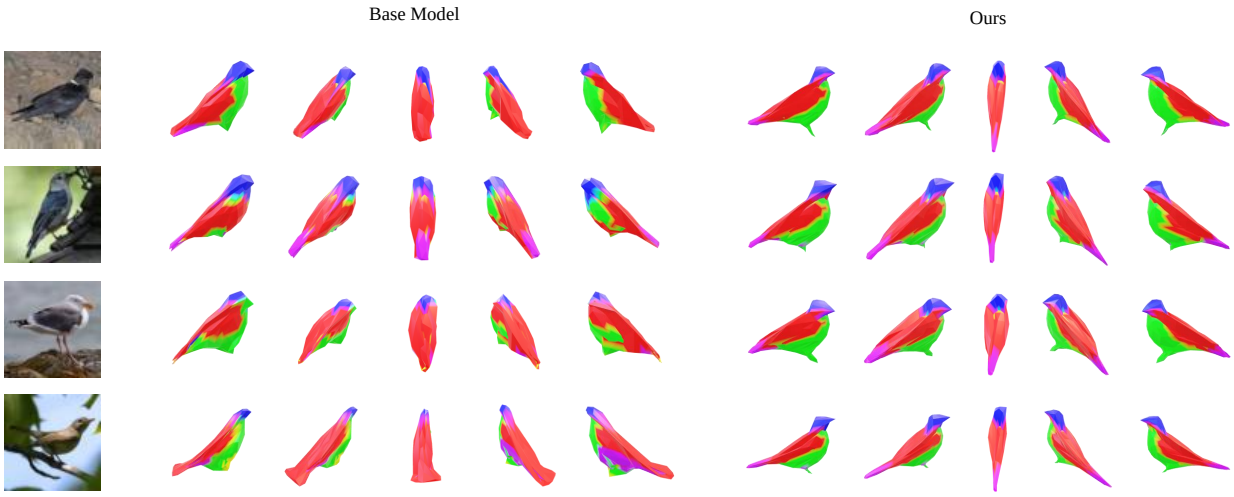

Figure 7. Visualization of partonomic reconstruction based on CUB-200-2011. Each row is a bird. The first column is the input image, the other columns are the partonomic reconstruction from base model and our proposed method with five different viewpoints.

ing mesh vertices. We render the images and object masks with the original meshes in ShapeNet and render part masks using the part-segmented meshes, shown in Figure 2.

2.3. CUB-200-2011

Since the official dataset does not provide the detailed 2D part labels, we use part masks provided by [1], which only labels the first 70 categories. In addition, we merge the original 11 parts into 6 parts to avoid too small parts such bird eyes. Then, we label different parts with different colors, head (blue), body (green), wing (red), tail (purple), neck (cyan), and legs (yellow). The examples are visualized in Figure 3.

3. More Qualitative Results

We present more qualitative results on ShapeNetPart (Sec. 3.1), PartNet (Sec. 3.2), and CUB-200-2011 (Sec. 3.3).

3.1. ShapeNetPart

Figure 4 visualizes the qualitative results across five categories in ShapeNetPart based on three approaches. We extend Unicorn [10] and AST [7] by modeling the part class of each mesh vertex, and adding the part rendering loss for learning. These two extension are denoted as Unicorn* and AST*. As shown in Figure 4, our proposed method is able to obtain both better overall and partonomic reconstructions. For the airplane category, the reconstruction generated by our proposed method has more clear-boundary wings compared with the output of the other two approaches.

3.2. PartNet

Figure 5 visualizes the qualitative results of three approaches in PartNet. Each column represents an object, while each row represents the input image, the ground truth, and the reconstruction generated by Unicorn*, AST* and our proposed method, respectively. The visualization highlights our proposed method outperforms Unicorn* and AST* across all five categories in terms of overall and partonomic reconstruction. Specifically, for the mug category, Unicorn* and AST* show artifacts in the handle region and fail to achieve consistent part segmentations. However, our proposed method better captures the handle's geometry and the part decomposition, achieving a more realistic reconstruction.

3.3. CUB-200-2011

Figure 6 demonstrates qualitative results illustrating the front and back views of the reconstructed 3D shapes on CUB-200-2011. These extended views offer a more comprehensive evaluation of the reconstruction quality.

Figure 7 shows the qualitative comparison between our proposed method and the base model on CUB-200-2011 [14]. We tried to train Unicorn* and AST* on this dataset with part rendering loss, but got unsatisfactory results. Each row of Figure 7 represents a bird. The first column is the input image, the second to sixth columns are the partonomic reconstruction of base model from different viewpoints, and the seventh to eleventh columns show our partonomic reconstruction from five different viewpoints. It shows our proposed method is able to generate the more accurate overall and semantic-meaningful partonomic reconstruction. Specifically, the reconstructed tail of the last in-

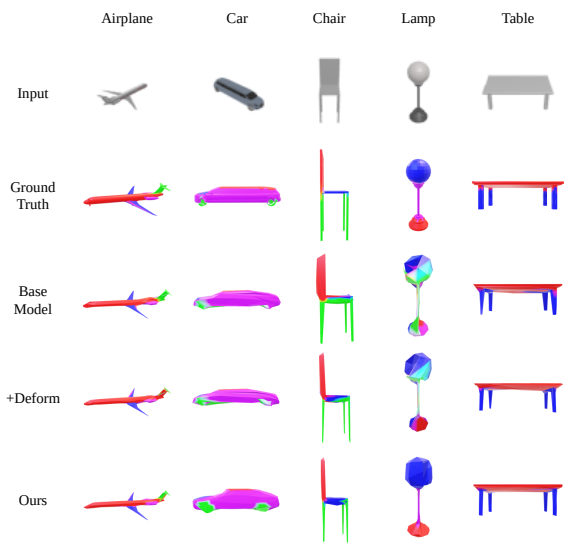

Figure 8. **More qualitative results on ShapeNetPart.**

Sup.	Method	Average	Airplane	Car	Chair	Lamp	Table
2D	Unicorn	0.249	0.099	0.157	0.243	0.499	0.247
	AST	0.217	0.090	0.151	0.222	0.393	0.229
	Ours	**0.197**	**0.082**	**0.148**	**0.227**	**0.340**	**0.189**
3D	Unicorn	0.138	0.070	0.138	0.151	0.183	0.147
	AST	0.132	0.068	0.134	0.142	0.178	0.140
	Ours	**0.127**	**0.059**	**0.129**	**0.139**	**0.176**	**0.131**

Table 3. **Quantitative comparison using 2D and 3D supervisions on ShapeNetPart.** The Chamfer-L_1 performance is reported. 2D supervision means object and part masks. 3D supervision means ground truth meshes.

stance is coarser than ours. In addition, the part labels are mixed and more messy.

4. Ablation Study

Component analysis. Figure 8 provides more qualitative results on ShapeNetPart [4]. Across all five categories, our proposed method reconstructs the 3D shapes which are much closer to the ground truth compared to the Base model and +Deform, with more clear boundary segmentation.

2D versus 3D supervisions. Table 3 shows quantitative comparison between 2D object and part mask supervisions and 3D mesh supervisions on ShapeNetPart [4]. Our proposed method consistently outperforms Unicorn [10] and AST [7] across all five categories. It is worth noting that reconstruction performance under full 3D supervision is not comparable to that of prior works [6, 12] because the ShapeNetPart used in this paper is significantly smaller than the full ShapeNet.

5. Limitation

Since our compositional 3D object representation is defined on a spherical mesh, it struggles to effectively handle high-genus shape topologies, such as kettles and donuts. For future work, we plan to extend our compositional approach to implicit representations, such as the signed distance function, which can represent shapes with arbitrary topologies.

References

[1] Hamed Behzadi-Khormouji and José Oramas. A protocol for evaluating model interpretation methods from visual explanations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1421–1429, 2023. 4

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[3] Mathieu Desbrun, Mark Meyer, Peter Schröder, and Alan H Barr. Implicit fairing of irregular meshes using diffusion and curvature flow. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 317–324, 1999. 1

[4] Haoqiang Fan, Hao Su, and Leonidas 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*, pages 605–613, 2017. 1, 5

[5] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 88–104. Springer, 2020. 1

[6] Zixuan Huang, Stefan Stojanov, Anh Thai, Varun Jampani, and James M Rehg. Zeroshape: Regression-based zero-shot shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10061–10071, 2024. 5

[7] Dian Jia, Xiaoqian Ruan, Kun Xia, Zhiming Zou, Le Wang, and Wei Tang. Analysis-by-synthesis transformer for single-view 3d reconstruction. In *European Conference on Computer Vision*, pages 259–277, 2024. 1, 3, 4, 5

[8] Charles Loop. Smooth subdivision surfaces based on triangles. 1987. 1, 3

[9] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019. 2

[10] Tom Monnier, Matthew Fisher, Alexei A Efros, and Mathieu Aubry. Share with thy neighbors: Single-view reconstruction by cross-instance consistency. In *European Conference on Computer Vision*, pages 285–303. Springer, 2022. 1, 3, 4, 5

[11] Andrew Nealen, Takeo Igarashi, Olga Sorkine, and Marc Alexa. Laplacian mesh optimization. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pages 381–389, 2006. 1

[12] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9964–9973, 2019. 5

[13] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[14] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 4

[15] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1