Supplementary Material of KP-RED: Exploiting Semantic Keypoints for Joint 3D Shape Retrieval and Deformation

Ruida Zhang[1*], Chenyangguang Zhang[1*], Yan Di[2], Fabian Manhardt[3],
Xingyu Liu[1], Federico Tombari[2,3], Xiangyang Ji[1]
[1]Tsinghua University, [2]Technical University of Munich, [3]Google
{zhangrd23@mails, zcyg22@mails, xyji@}.tsinghua.edu.cn *

1. Network Architecture

The detailed network architecture of KP-RED is shown in Fig. 1.

For the **Retrieval** module, we first use the PointNet-based [10] keypoint predictor to discover the keypoints (Fig. 1 (a)). Then we use a 2-layer MLP to extract point-wise features. We aggregate the local features of each keypoint within its support region and adopt a YOGO-like [13] self-attention module to discover region-to-region relations (Fig. 1 (b)). We replace the Farthest Keypoint Sampling strategy in [13] with our semantic-consistent keypoint detection approach. We use the relation inference module [13] to extract the region tokens. Finally, we concatenate all local retrieval tokens of each region in the uniform order and yield the global token by concatenating them together.

During training, we adopt the auxiliary reconstruction task to supervise the learning of the **Retrieval** module. Given the region tokens and the target keypoints, the reconstruction network generates the corresponding region of the deformed shape. We randomly sample $N_S = 192$ vectors $P^* \in \mathbb{R}^3$ on a unit sphere and concatenate them with the region tokens and target keypoints together to serve as the input of the reconstruction network. The reconstruction network is required to project P^* from the unit sphere to the corresponding positions on the deformed shape indicated by the retrieval tokens. The reconstruction network is constructed with the 4-layer PointNet (Fig. 1 (c)).

For the **Deformation** module, we use the same keypoint predictor as the **Retrieval** module to detect keypoints (Fig. 1 (a)) and the same architecture for the self-attention module to extract region tokens (Fig. 1 (b)). We follow [7] and compute the cage with $N_C = 42$ vertices to control deformation. Given the extracted region token of each keypoint, we concatenate it with a one-hot encoding of the keypoint index and use a 3-layer MLP to predict the influence vector of the keypoint (Fig. 1 (d)).

*Authors with equal contributions.

2. Partial Shape Generation

Given an occlusion ratio, we augment the point cloud of the full shape to generate partial shape by random slicing. Given the point cloud $P \in \mathbb{R}^{3\times N}$ with N_P points and the occlusion ratio r_o, our objective is to find a plane $\mathbf{n}^T p = d$ with normal vector $\mathbf{n} \in \mathbb{R}^{3\times 1}$ and distance $d \in \mathbb{R}$, such that we can slice the point cloud P with this plane and remove all points on one side of the plane, so that $r_o N_P$ points are removed. We first randomly choose the plane normal $\mathbf{n}$. Then we adjust the position of the plane, such that the given ratio of points are removed after slicing. Specifically, the distance of the plane d can be derived from the following equation,

$$d = f_{top}(\mathbf{n}^T P, r_o N_P) \tag{1}$$

where $f_{top}(A, k)$ denotes the top k value in the vector A. All points that satisfy $\mathbf{n}^T p \geq d$ are removed.

3. Evaluation for Partial Shapes

The typical bilateral Chamfer Distance (CD) between two shapes S_1, S_2 is the sum of two Unilateral Chamfer Distance (UCD),

$$f_{CD}(S_1, S_2) = f_{UCD}(S_1, S_2) + f_{UCD}(S_2, S_1) \tag{2}$$

where the UCD is defined as,

$$f_{UCD}(S_1, S_2) = \frac{1}{|S_1|} \sum_{x\in S_1} \min_{y\in S_2} ||x - y||_2^2 \tag{3}$$

For each point in S_1, $f_{UCD}(S_1, S_2)$ finds the nearest point in S_2, and sums the square of distance up.

When handling partial input, the typical bilateral CD does not reflect the **R**&**D** quality well since some parts of the target shape S_{tgt} are missing and not every point in the deformed shape $S_{src2tgt}$ has corresponding point in S_{tgt}. Therefore, we use Unilateral Chamfer Distance $f_{UCD}(S_{src2tgt}, S_{tgt})$ as the evaluation metrics for partial shape reconstruction.

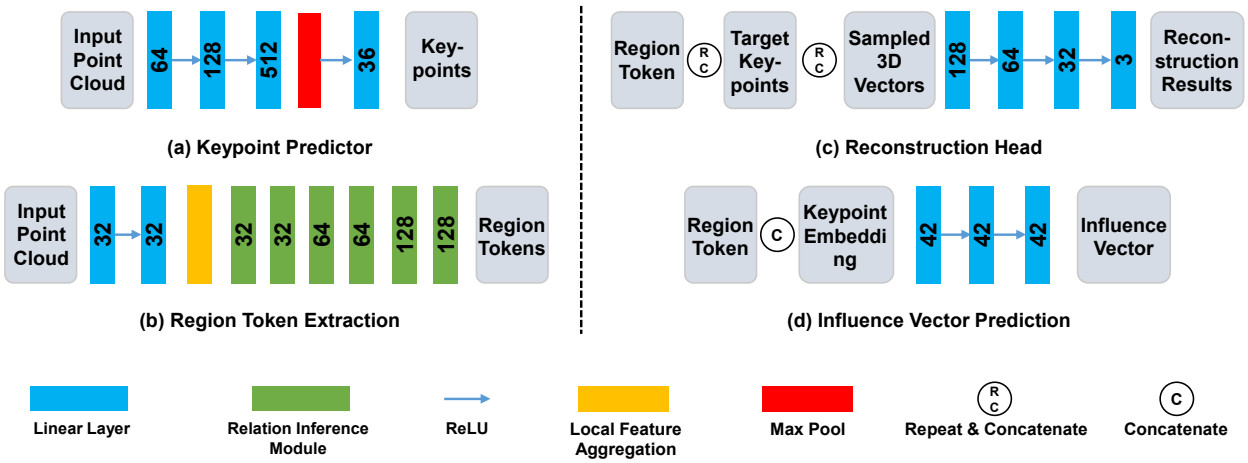

Figure 1. Network architecture of KP-RED.

Evaluation on Scan2CAD Dataset [1]. The target point cloud in Scan2CAD is noisy and inaccurate due to the limitations of the depth sensors and thus not suitable to calculate the evaluation metrics. Therefore, we use the ground truth models to generate the clean target point cloud and use it to calculate the Unilateral Chamfer Distance metrics. Specifically, we use the ground truth pose to render the object to obtain its ground truth depth and then back-project the depth to obtain the clean point cloud.

4. Results with Top-K Retrieval Candidates

We further provide top-1, 5, 10, 25 results on 4 datasets in Tab. 1. Our method surpasses all competitors by a large margin under all evalution metrics.

PartNet				
Method	Top-1	Top-5	Top-10	Top-25
U-RED [5]	0.658	0.609	0.551	0.488
Uy *et al.* [11]	0.726	0.688	0.637	0.535
ShapeFlow [8]	0.452	0.403	0.340	0.247
Ours	**0.185**	**0.102**	**0.089**	**0.084**
PartNet 25% Occlusion				
U-RED [5]	0.226	0.203	0.180	0.172
Uy *et al.* [11]	0.214	0.199	0.179	0.165
Ours	**0.085**	**0.059**	**0.049**	**0.049**
PartNet 50% Occlusion				
U-RED [5]	0.278	0.265	0.255	0.228
Uy *et al.* [11]	0.328	0.267	0.241	0.218
Ours	**0.103**	**0.082**	**0.056**	**0.048**
Scan2CAD				
U-RED [5]	0.316	0.249	0.207	0.183
Uy *et al.* [11]	0.293	0.255	0.210	0.190
Ours	**0.097**	**0.081**	**0.060**	**0.052**

Table 1. Average (Unilateral) Chamfer Distance with top-K retrieval. Overall best results are **in bold**.

5. Oracle Retrieval Experiments

In Tab. 2, we perform an oracle retrieval experiment by deforming all the source shapes and choosing the one with minimum Chamfer Distance error. Our deformation module achieves the lowest Chamfer Distance error with respect to [11] and [8].

Method	Chair	Table	Cabinet	Average
Uy *et al.* [11]	0.643	0.564	0.494	0.592
Uy *et al.* w/ IDO [11]	0.583	0.482	0.494	0.526
ShapeFlow [8]	0.167	0.223	0.353	0.208
Ours	**0.089**	**0.079**	**0.097**	**0.084**

Table 2. Chamfer Distance metrics for **deformation** module using oracle retrieval. Overall best results are **in bold**.

6. Definitions of Loss Terms

Full shapes. We adopt two loss terms to train the deformation module with full shapes. The first one is the similarity loss $\mathcal{L}_{sim}$. Given the target shape S_{tgt} and the deformed shape $S_{src2tgt}$, we define the similarity loss as

$$\mathcal{L}_{sim} = f_{CD}(S_{src2tgt}, S_{tgt}), \tag{4}$$

where f_{CD} denotes the Chamfer Distance.

The second one is the keypoint regularization loss $\mathcal{L}_{kpt}$. Given the predicted keypoints of the source shape $\mathbf{K}_{src}$, we use Farthest Point Sampling to sample N_K points P_{fps} on the source shape. The regularization loss is defined as,

$$\mathcal{L}_{kpt} = f_{CD}(\mathbf{K}_{src}, P_{fps}). \tag{5}$$

Partial shapes. When handling the partial shapes, we train the parameters of keypoint predictor with two loss terms. The first one is the unilateral similarity loss $\mathcal{L}_{usim}$,

$$\mathcal{L}_{usim} = f_{UCD}(S_{src2tgt}, S_{tgt}), \tag{6}$$

where f_{UCD} denotes the Unilateral Chamfer Distance.

The second one is the weighted keypoint loss as introduced in the main text.

7. Limitations

- The effectiveness of **R**&**D** heavily relies on the availability of a pre-prepared model database consisting of considerable amounts of high-quality CAD models. Despite the existence of large-scale CAD model datasets such as ShapeNet [2], there are still some categories that remain uncovered. Moreover, it is essential to ensure that the source shapes in the database are representative and cover the shape variation across each category as much as possible.
- Currently the performance of KP-RED is highly influenced by the accuracy of pose from off-the-shelf pose estimation methods [6, 15, 16]. However, since it is hard to guarantee good generalization ability of these methods in numerous real-world scenarios, the performance of **R**&**D** will also deteriorate due to the failure of pose estimation. One possible solution is to integrate pose estimation into the **R**&**D** pipeline via vector neuron network [3, 4] where SO(3) invariant features can be extracted for downstream tasks, eliminating the influence of pose.
- While the cage-based deformation approach yields high-quality results and preserves fine-grained geometric details, it is theoretically unable to generate a completely perfect match to the target shape. As a result, the theoretical upper bound of our performance is limited. To overcome this limitation, additional neural techniques [12, 14] can be utilized to refine the **R**&**D** results further.

8. Additional Qualitative Results

Fig. 2 visualizes detected keypoints on PartNet [9], which demonstrates the semantic consistency of the detected keypoints. Fig. 3 shows **failure cases** on PartNet. Many failure cases result from shape variations that are not encompassed within the model database, such as a chair featuring a separate footrest or a table with an open drawer depicted in Fig. 3. Fig. 4 shows additional qualitative results on Scan2CAD [1]. KP-RED is much more robust in the real-world scenario and yield more precise **R**&**D** results than the competitors. Fig. 5 displays visualization for full shapes on PartNet [9]. Fig. 6, Fig. 7 and Fig. 8 exhibits qualitative comparisons for partial shapes under different occlusion levels on augmented PartNet. The results reveal our KP-RED performs robustly under different occlusion scenarios.

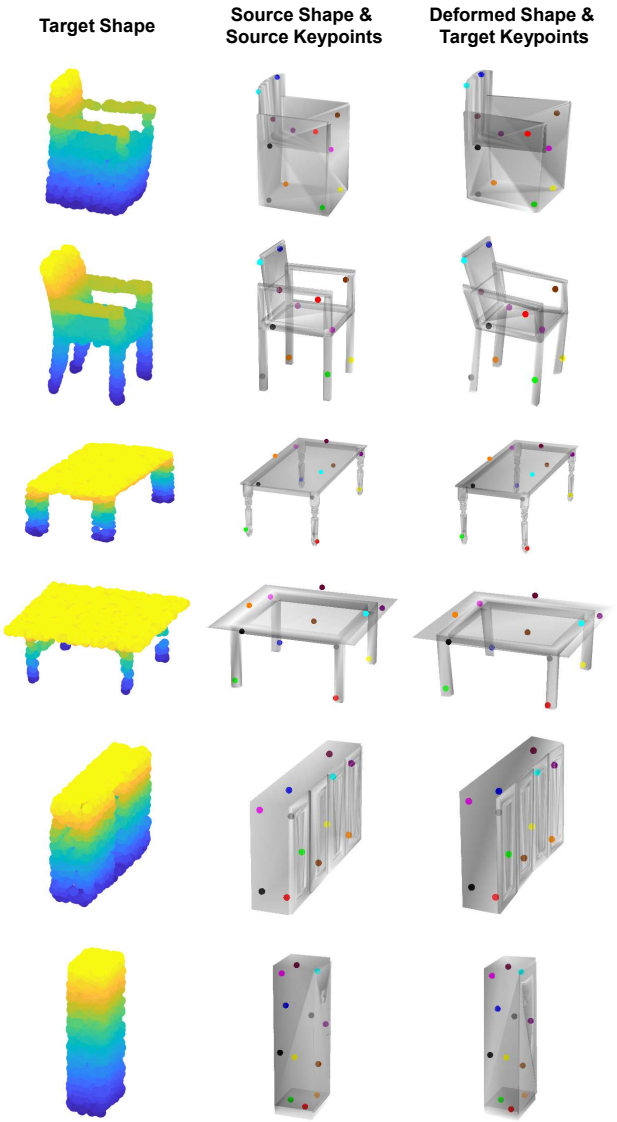

Figure 2. Visualization of detected keypoints on PartNet [9]. We use different colors for different keypoints to show their semantic consistency. The **R**&**D** results are rendered on the RGB images for better visualization.

References

[1] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X Chang, and Matthias Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2614–2623, 2019. 2, 3, 5

[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:

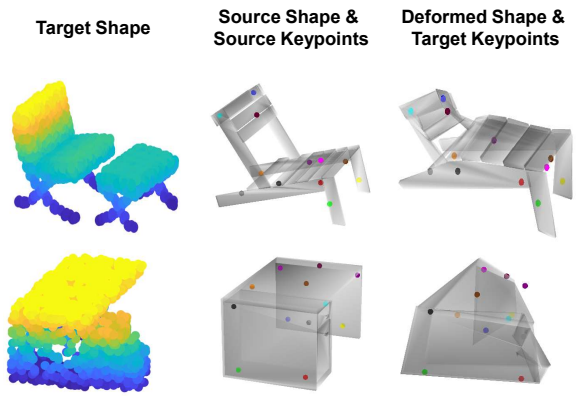

Figure 3. **Failure cases** on PartNet [9]. We use different colors for different keypoints to show their semantic consistency. The **R**&**D** results are rendered on the RGB images for better visualization.

An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[3] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12200–12209, 2021. 3

[4] Yan Di, Chenyangguang Zhang, Chaowei Wang, Ruida Zhang, Guangyao Zhai, Yanyan Li, Bowen Fu, Xiangyang Ji, and Shan Gao. Shapemaker: Self-supervised joint shape canonicalization, segmentation, retrieval and deformation. *arXiv preprint arXiv:2311.11106*, 2023. 3

[5] Yan Di, Chenyangguang Zhang, Ruida Zhang, Fabian Manhardt, Yongzhi Su, Jason Rambach, Didier Stricker, Xiangyang Ji, and Federico Tombari. U-red: Unsupervised 3d shape retrieval and deformation for partial point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8884–8895, 2023. 2

[6] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6781–6791, 2022. 3

[7] Tomas Jakab, Richard Tucker, Ameesh Makadia, Jiajun Wu, Noah Snavely, and Angjoo Kanazawa. Keypointdeformer: Unsupervised 3d keypoint discovery for shape control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12783–12792, 2021. 1

[8] Chiyu Jiang, Jingwei Huang, Andrea Tagliasacchi, and Leonidas J Guibas. Shapeflow: Learnable deformation flows among 3d shapes. *Advances in Neural Information Processing Systems*, 33:9745–9757, 2020. 2

[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. 3, 4

[10] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1

[11] Mikaela Angelina Uy, Vladimir G Kim, Minhyuk Sung, Noam Aigerman, Siddhartha Chaudhuri, and Leonidas J Guibas. Joint learning of 3d shape retrieval and deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11713–11722, 2021. 2

[12] Zhen Xing, Yijiang Chen, Zhixin Ling, Xiangdong Zhou, and Yu Xiang. Few-shot single-view 3d reconstruction with memory prior contrastive network. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 55–70. Springer, 2022. 3

[13] Chenfeng Xu, Bohan Zhai, Bichen Wu, Tian Li, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. You only group once: Efficient point-cloud processing with token representation and relation inference module. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4589–4596. IEEE, 2021. 1

[14] Shuo Yang, Min Xu, Haozhe Xie, Stuart Perry, and Jiahao Xia. Single-view 3d object reconstruction from shape priors in memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3152–3161, 2021. 3

[15] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. In *European Conference on Computer Vision*, pages 655–672. Springer, 2022. 3

[16] Ruida Zhang, Yan Di, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Ssp-pose: Symmetry-aware shape prior deformation for direct category-level object pose estimation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7452–7459. IEEE, 2022. 3

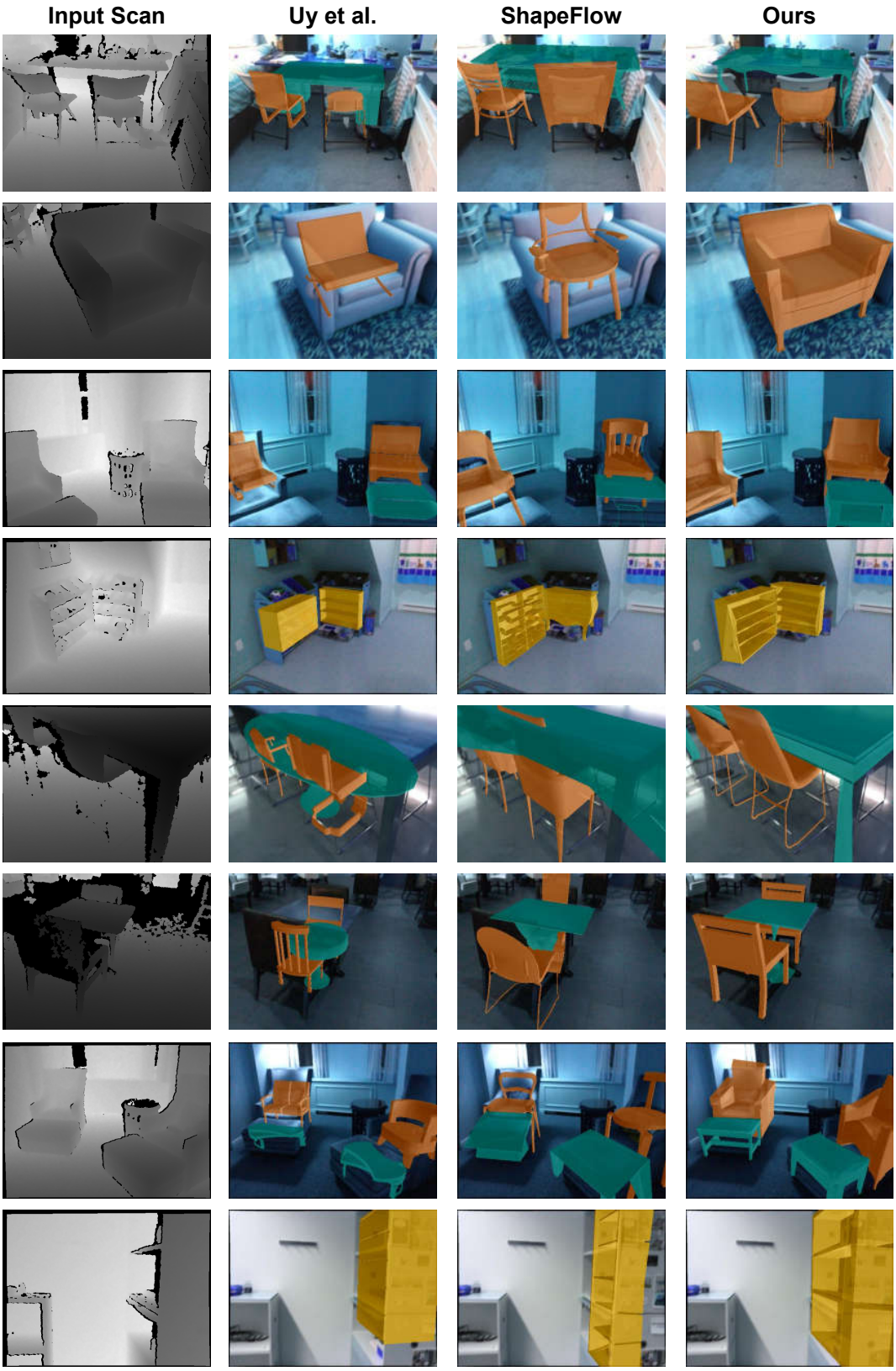

Figure 4. Additional qualitative results on Scan2CAD [1]. The **R**&**D** results are rendered on the RGB images for better visualization.

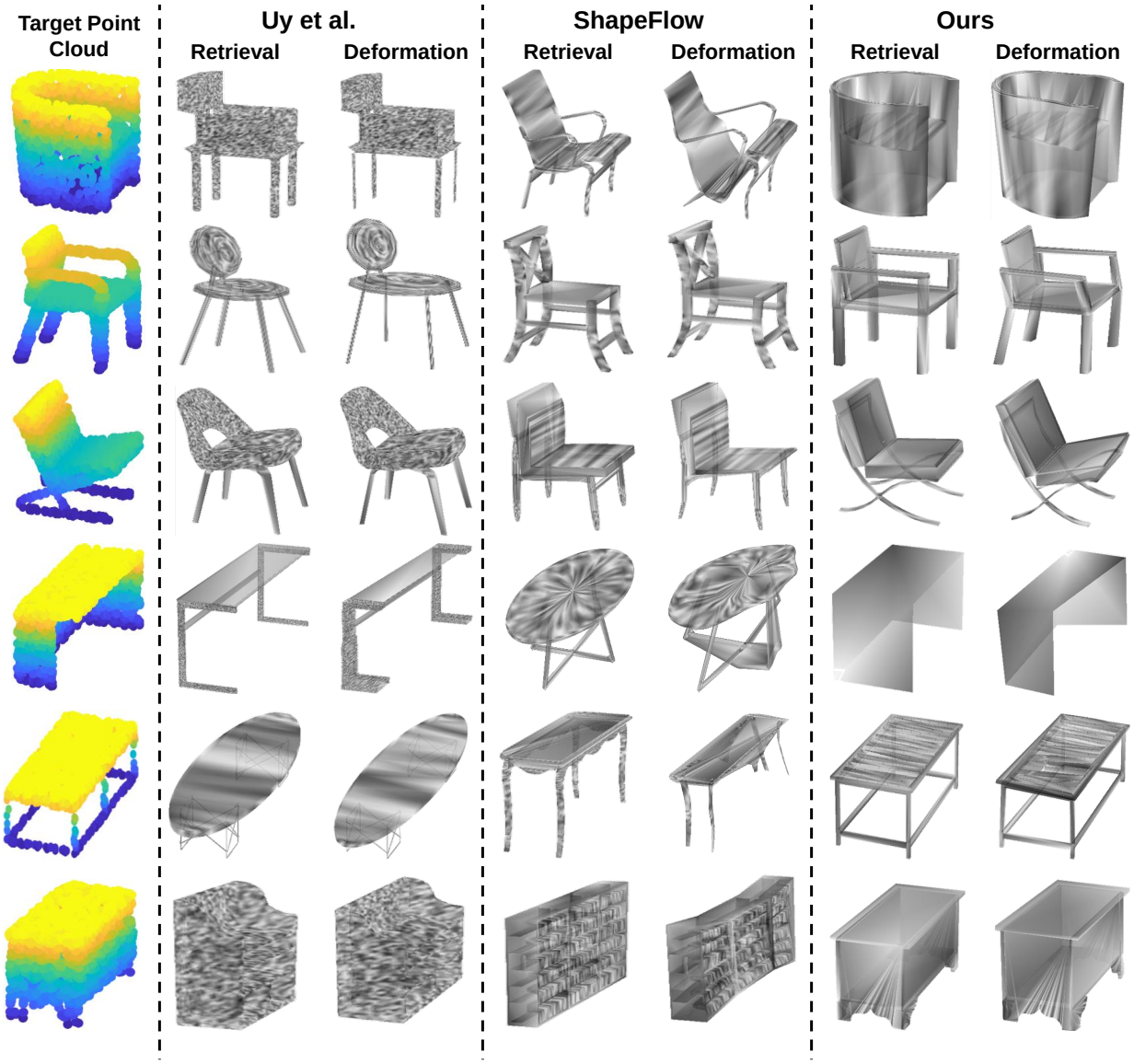

Figure 5. Additional qualitative results for full shapes on PartNet.

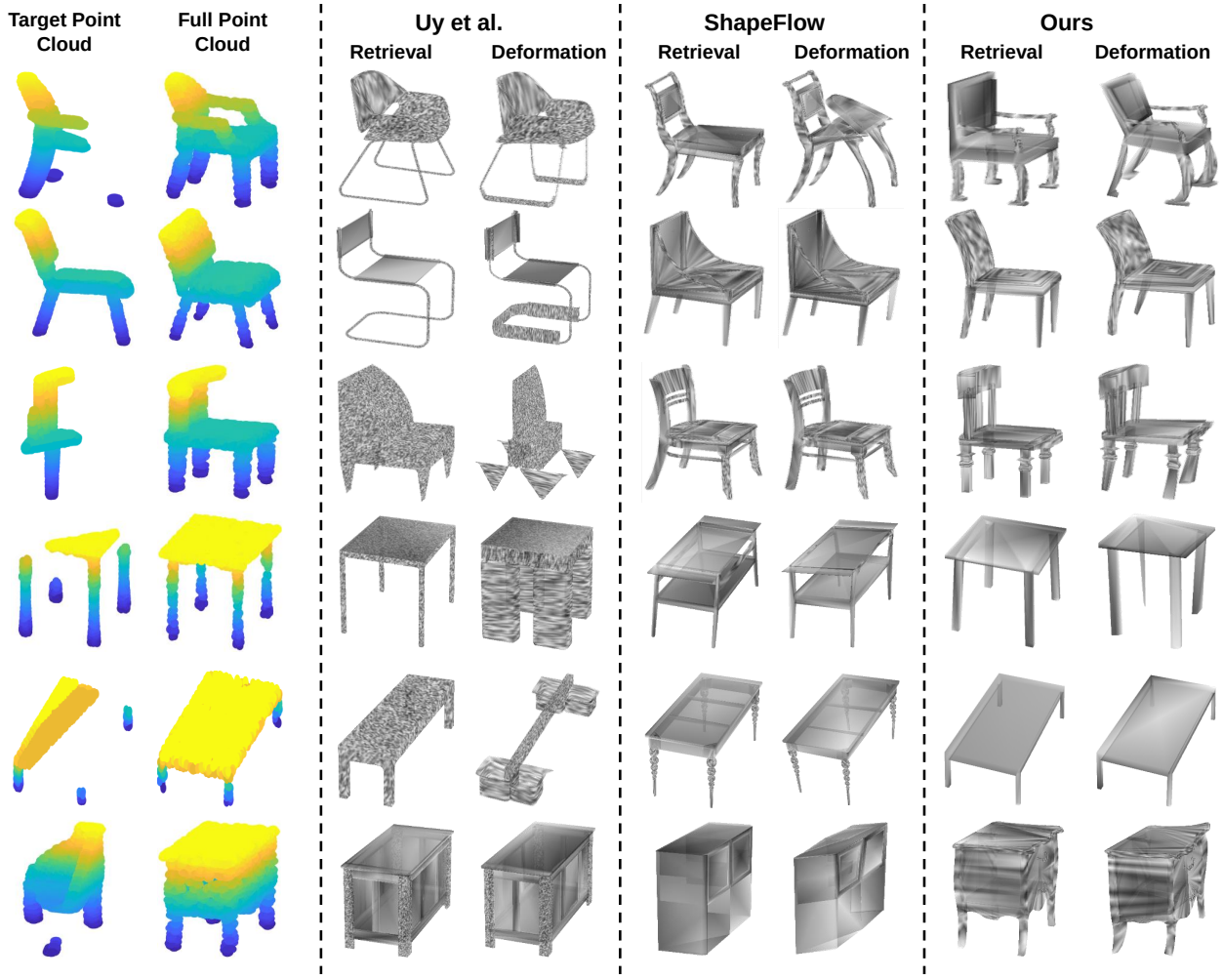

Figure 6. Additional qualitative results for partial shapes with the occlusion ratio of 75% on augmented PartNet.

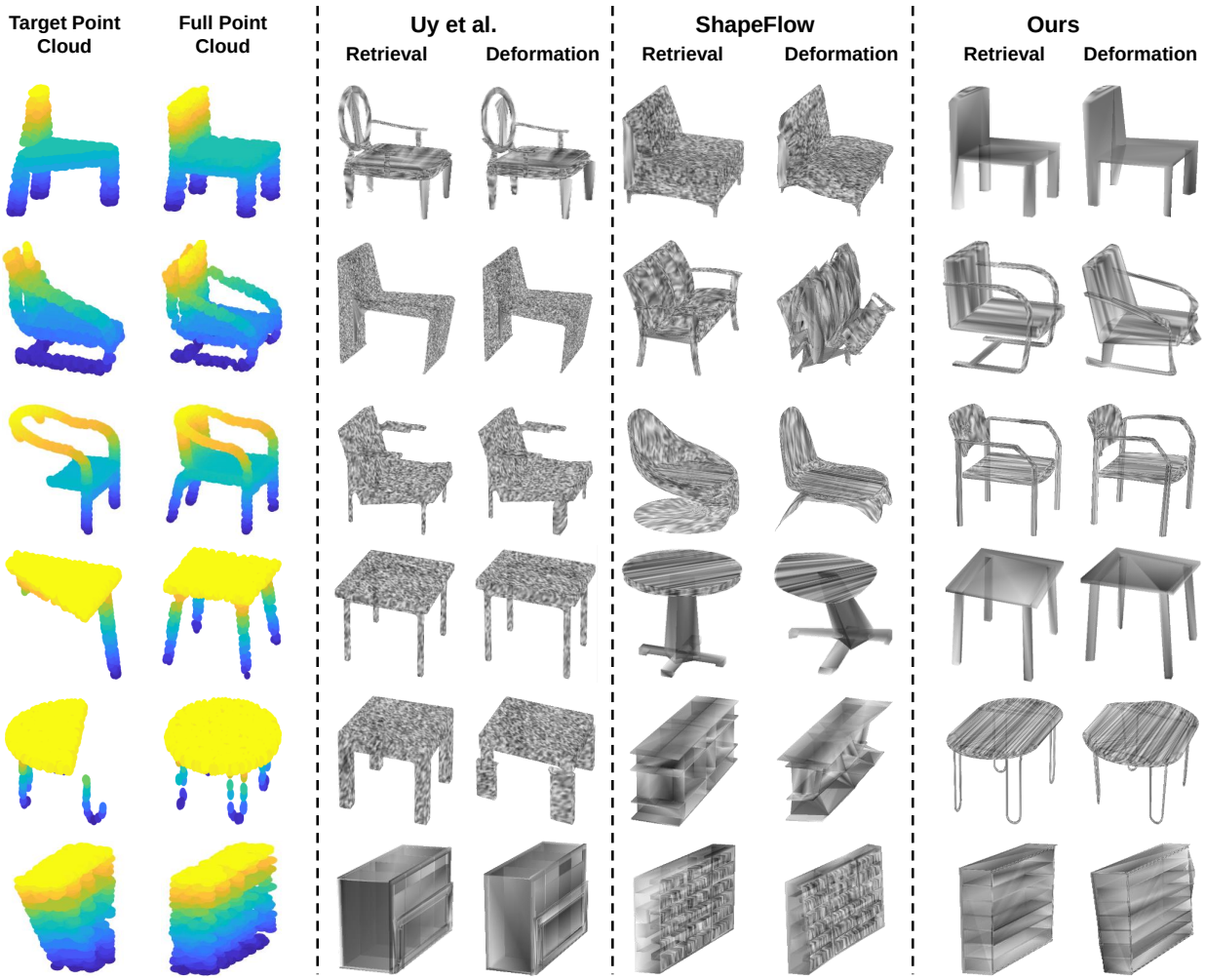

Figure 7. Additional qualitative results for partial shapes with the occlusion ratio of 50% on augmented PartNet.

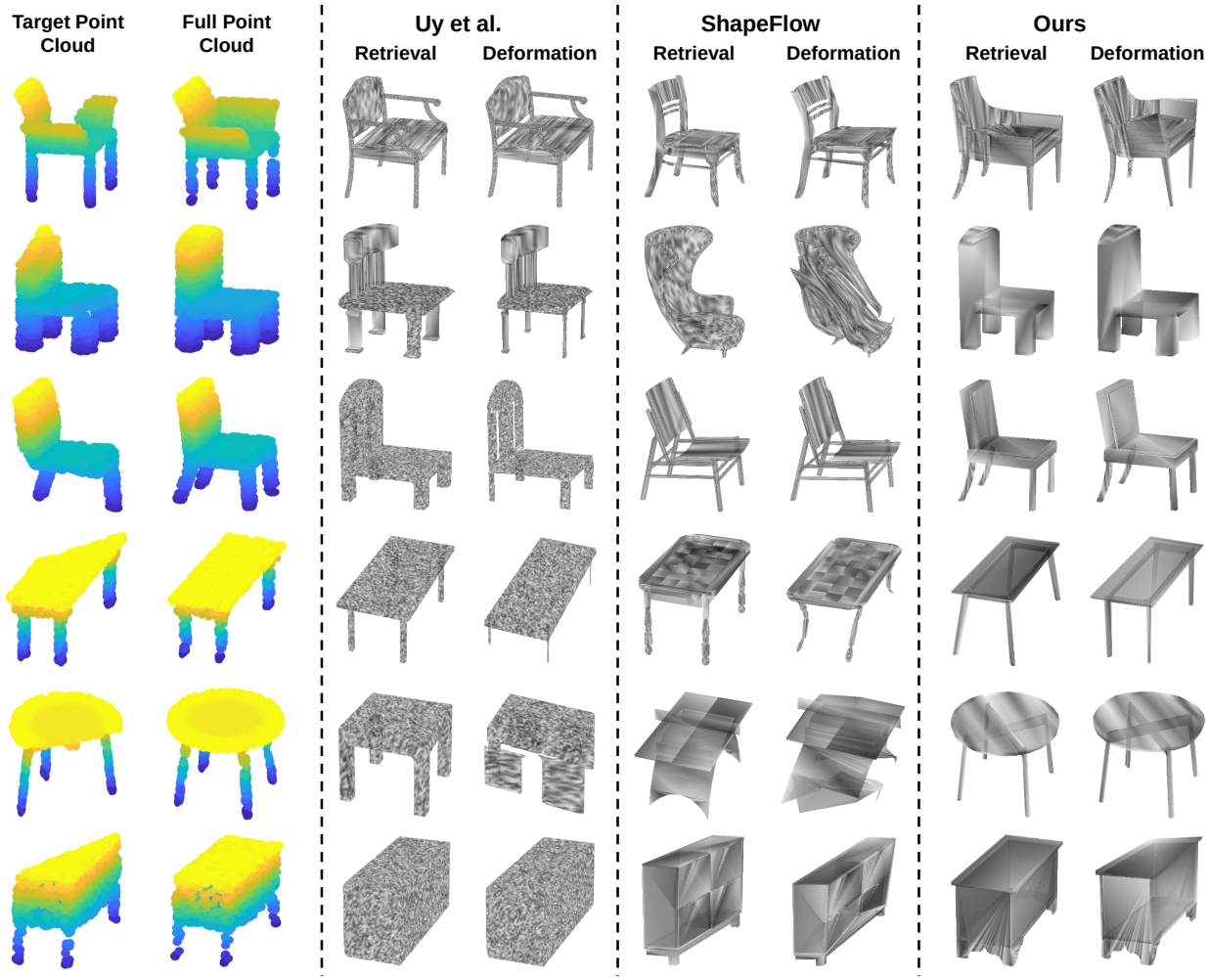

Figure 8. Additional qualitative results for partial shapes with the occlusion ratio of 25% on augmented PartNet.