Supplementary Material of U-RED: Unsupervised 3D Shape Retrieval and Deformation for Partial Point Clouds

Yan Di[1*], Chenyangguang Zhang[2*], Ruida Zhang[2*], Fabian Manhardt[3], Yongzhi Su[4], Jason Rambach[4], Didier Stricker[4], Xiangyang Ji[2] and Federico Tombari[1,3]
[1]Technical University of Munich, [2]Tsinghua University, [3] Google, [4] DFKI
{yan.di@, tombari@in.}tum.de, {zcyg22, zhangrd21}@mails.tsinghua.edu.cn

1. Network Architecture

Fig. 1 illustrates the detailed network architecture that we employed for U-RED, including parallel feature encoders (a), residual-guided retrieval network (b), graph attention based deformation network (c) and reconstruction head (d). Note that the reconstruction head is only used for joint training. The input point cloud is roughly calibrated to the canonical view by using arbitrary category-level pose estimators [2, 3, 7, 8]. Since the network itself has a certain resistance to rotational interference, we don't need the estimated rotation results to be extremely accurate. In this paper, for fair comparisons, we follow ROCA [3]'s pose estimator.

The parallel feature encoders are adopted for extracting point-wise and global features $\{\mathcal{F}^p, \mathcal{G}^p\}$ from the target partial point cloud $\mathcal{T}^p$, $\{\mathcal{F}^f, \mathcal{G}^f\}$ for the corresponding full shape $\mathcal{T}^f$ (only in training) and $\{\mathcal{F}^d, \mathcal{G}^d\}$ for sources shapes $\mathcal{O}^c$ in the database, as introduced in Sec. 3.1 of our main text. All encoders for $\mathcal{T}^p$, $\mathcal{T}^f$ and $\mathcal{O}^c$ share the same network architecture. The point-wise features are extracted by the MLP directly, while the corresponding global features are generated by applying max-pooling to the aforementioned point-wise features.

Our residual-guided retrieval network is a 4-layers MLP with batch normalization [4]. The input of the retrieval network is the concatenation of partial features $\{\mathcal{F}^p, \mathcal{G}^p\}$, source shape features $\mathcal{G}^d$ and the normalized full shape indicator $\hat{\mathcal{G}}^f$. In training, $\hat{\mathcal{G}}^f$ is the normalized $\mathcal{G}^f$. During inference, $\hat{\mathcal{G}}^f$ is replaced by $\mathcal{G}^s$ which is sampled on the surface of the unit sphere Ω. Taking these features, the retrieval network outputs the residual field $R = \{R_i \in \mathbb{R}^3, i = 1, ..., M\}$ to accomplish our noise-robust *one-to-many* retrieval.

The deformation network utilizes an AGNN with cross-attention and self-attention module as described in Sec. 3.3 of the main text. After AGNN, the updated part features $\mathcal{P}^f$ are fed to an MLP to predict center displacement $\mathcal{C}_d$ and axis-aligned scaling parameters $\{s_w, s_h, s_l\}$ to obtain the bounding box parameters of each part.

The reconstruction head is a supplementary branch used only in the training stage, which takes the concatenated features $\{\mathcal{F}^*, \mathcal{G}^*\}$ as input and reconstructs the input point cloud. We stack three parallel reconstruction heads to respectively process the target partial point cloud $\mathcal{T}^p$, the target full point cloud $\mathcal{T}^f$ and the source shape $\mathcal{O}^c$. The reconstruction heads are utilized to help the encoder extract representative features. The used architecture is again an MLP network with Batch Normalization layers.

2. Additional Loss Terms

The basic loss term $\mathcal{L}^b$ introduced in Sec. 3.4 of our main manuscript includes a Chamfer Distance loss $\mathcal{L}^{cd}$, a reconstruction loss $\mathcal{L}^r$, and an optional symmetry loss $\mathcal{L}^{sym}$.

The Chamfer Distance loss $\mathcal{L}^{cd}$ [6] and the symmetry loss $\mathcal{L}^{sym}$ are utilized mainly for supervising the deformation process. $\mathcal{L}^{cd}$ is calculated by the predicted deformed shape $\tilde{\mathcal{O}}^c$ and the input target $\mathcal{T}^p$. All models in our database have reflective symmetry [6]. The yz-plane of our source shapes is aligned with the symmetry axis. Thereby, given the predicted deformed shape $\tilde{\mathcal{O}}^c$, we flip $\tilde{\mathcal{O}}^c$ with respect to the yz-plane and obtain $\tilde{\mathcal{O}}'$, thus we obtain $\mathcal{L}^{sym}$ as

$$\mathcal{L}^{sym} = \mathcal{L}^{cd}(\tilde{\mathcal{O}}^c, \tilde{\mathcal{O}}') \tag{1}$$

Note that Uy *et al.* [6] has demonstrated that $\mathcal{L}^{sym}$, serving as the regularization, can enforce bilateral symmetry of the output deformed shapes, leading to more geometrically precise deformation.

We propose the reconstruction loss $\mathcal{L}^r$ for the reconstruction heads described in Sec 1. Given the input shape $\mathcal{T}$, the corresponding reconstruction head predicts the recon-

*Authors with equal contributions.

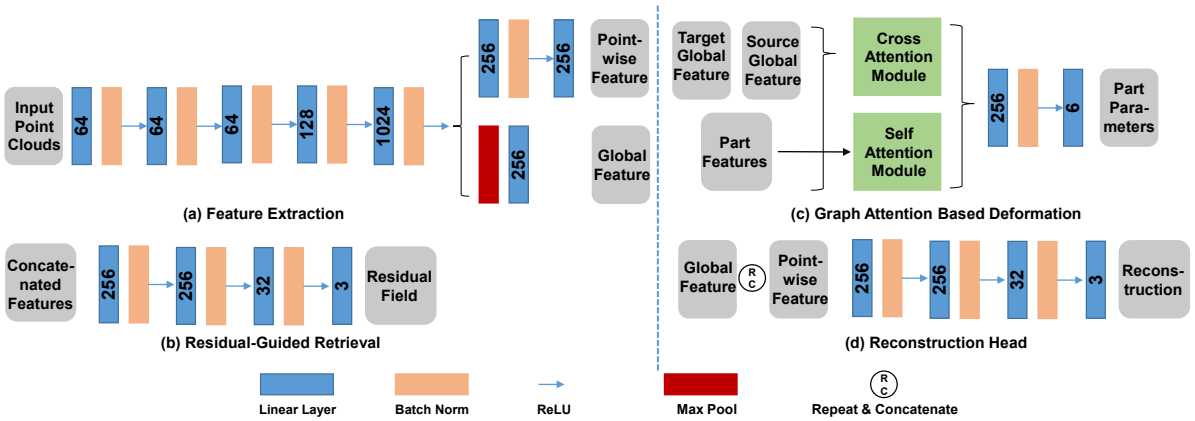

Figure 1. Network architecture of U-RED.

structed shape $\mathcal{T}'$. We calculated $\mathcal{L}^r$ according to

$$\mathcal{L}^r = \|\mathcal{T} - \mathcal{T}'\|^2, \tag{2}$$

where we use the $\mathcal{L}_2$ Euclidean Distance between point sets as loss function. Such $\mathcal{L}^r$ is utilized for all reconstruction heads for $\mathcal{T}^p$, $\mathcal{T}^f$ and $\mathcal{O}^c$.

3. Partial Data Generation

We generate our partial point clouds input $\mathcal{T}^p \in \mathbb{R}^{1024\times3}$ used in the training stage by simulating real-world partial observations, including random cropping, noise addition and random rotation. Given an available full CAD shape from the synthetic dataset [5], we first uniformly sample the point cloud $\mathcal{T} \in \mathbb{R}^{2048\times3}$, and then randomly crop 50% of the full point cloud to generate $\mathcal{T}_\nabla{}^p$. Next, we add zero-mean Gaussian noise having a variance of 0.05 to the coordinates of $\mathcal{T}_\nabla{}^p$ via $\mathcal{T}^p = (1+\sigma)\mathcal{T}_\nabla{}^p$. Finally, we apply a random rotation to $\mathcal{T}^p$, where the rotation angle along each axis is sampled from $\mathcal{N}(0, 5.0)$. The full shape input $\mathcal{T}^f \in \mathbb{R}^{1024\times3}$is generated by uniformly down-sampling $\mathcal{T}$.

4. Discussion on OTM module

Fig. 2 demonstrates the unique retrieval results of target objects by our one-to-many strategy. Since each partial shape may correspond to multiple full shapes, which further enables multiple possible retrievals, our one-to-many strategy consistently outperforms Uy *et al.* [6].

5. Effects of Noise Level.

As shown in Tab. 1, our method can work robustly under different noise levels. Note that we don't add very heavy noise in this experiment since too much noise will make the target object incompatible with the provided ground truth.

	max(R) / mean(R)			
Noise Std	Chair	Table	Cabinet	Average
0.05	1.02 / 0.95	1.95 / 1.33	1.61 / 1.30	1.53 / 1.17
0.1	1.77 / 2.43	2.63 / 1.88	2.03 / 2.06	2.22 / 2.13

Table 1. Results of different retrieval metrics under different input noise. We conduct experiments on PartNet and add different levels of Gaussian noise.

Methods	Chair	Table	Cabinet	Average
Uy *et al.* [6]	0.76	0.70	0.72	0.73
Ours	0.76	0.72	0.84	0.75

Table 2. Results of full shape input. We conduct experiments on PartNet.

Sample Times	Chair	Table	Cabinet
1	1.07	1.53	1.85
100	0.95	1.34	1.30
1000	0.95	1.33	1.30

Table 3. Ablation of sample times on PartNet.

6. Different Retrieval Metrics.

As shown in Tab. 1, min *mean(R)* works slightly better than min *max(R)* given the same residual field R.

7. Full Shape Results

For our method, we directly use the full shape feature extractor in the OTM module to conduct this experiment. As shown in Tab. 2, our method performs on par with Uy *et al.* given full shape as input.

Figure 2. Visualization of our *one-to-many* (OTM) retrieval technique on Scan2CAD [1] dataset. Our proposed residual-guided retrieval learns a *one-to-many* relationship to solve the problem that one partial shape may correspond to multiple full shapes. On the contrast, Uy *et al.* [6] takes *one-to-one* retrieval (OTO) which yields inferior performance facing with noisy and partial targets in real-world scenes.

8. Ablation of Sample Times

We sample 1000 times in the main paper to generate 1000 possible retrievals. We ablate the sample times in Tab. 3.

9. Additional Qualitative Results

9.1. Additional Qualitative Comparison

We provide additional qualitative comparison in both real-world (Fig. 4) and synthetic scenes (Fig. 3) with [6]. The results show that U-RED consistently outperforms Uy *et al.* [6] in both scenarios.

9.2. Failure cases

In the last two rows of Fig. 7 and the last row of Fig. 8 in the main manuscript, we additionally illustrate 3 failure cases. The main factor that influences the accuracy of retrieval and deformation lies in the quality of object detection and depth estimation. For heavily occluded objects, the detection and depth estimation results are typically erroneous, leading to bad retrieval results and thus inaccurate deformation.

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. 3, 4

[2] 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. 1

[3] Can Gümeli, Angela Dai, and Matthias Nießner. Roca: Robust cad model retrieval and alignment from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4022–4031, 2022. 1

[4] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 1

[5] 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, 4

[6] 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. 1, 2, 3, 4

[7] 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

Figure 3. Visualization on PartNet [5]. Qualitative comparison with Uy *et al.* [6] demonstrates that our U-RED performs more robustly, gains more accurate retrieval and more precise deformation results when facing partially observed point clouds.

European Conference on Computer Vision, pages 655–672. Springer, 2022. 1

[8] 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. 1

Figure 4. Qualitative results on Scan2CAD [1] dataset. U-RED consistently outperforms state-of-the-art Uy *et al.* [6].