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

Abstract

*In this paper, we propose **U-RED**, an **U**nsupervised shape **RE**trieval and **D**eformation pipeline that takes an arbitrary object observation as input, typically captured by RGB images or scans, and jointly retrieves and deforms the geometrically similar CAD models from a pre-established database to tightly match the target. Considering existing methods typically fail to handle noisy partial observations, U-RED is designed to address this issue from two aspects. First, since one partial shape may correspond to multiple potential full shapes, the retrieval method must allow such an ambiguous one-to-many relationship. Thereby U-RED learns to project all possible full shapes of a partial target onto the surface of a unit sphere. Then during inference, each sampling on the sphere will yield a feasible retrieval. Second, since real-world partial observations usually contain noticeable noise, a reliable learned metric that measures the similarity between shapes is necessary for stable retrieval. In U-RED, we design a novel point-wise residual-guided metric that allows noise-robust comparison. Extensive experiments on the synthetic datasets PartNet, ComplementMe and the real-world dataset Scan2CAD demonstrate that U-RED surpasses existing state-of-the-art approaches by 47.3%, 16.7% and 31.6% respectively under Chamfer Distance.*

1. Introduction

3D semantic scene perception [36, 56] involves the decomposition of a scene into its constituent objects, understanding and reconstructing all the detected objects, and putting them in place to formulate a holistic scene representation. Significant progress has been made recently in attaining the comprehensive analysis of multiple objects' geometry [17, 30, 42, 57, 61], dynamic reconstruction of both scene and objects [11, 12], structure-aware scene completion [5], *etc*. These methods demonstrate promising results in overall reconstruction quality but typically fail in preserving fine-grained geometric structures. To address this problem, **Retrieval** and **Deformation** (**R**&**D**) methods [8, 25, 35, 40, 46, 47, 49, 54] are proposed. They leverage a pre-prepared 3D shape database (usually represented as CAD models) as prior knowledge and typically follow a two-stage scheme to generate a clean and compact scene representation. First, based on manually picked metrics, the most similar shape of the target is selected from the database. Then, the retrieved shape is scaled, aligned and rotated to match the target.

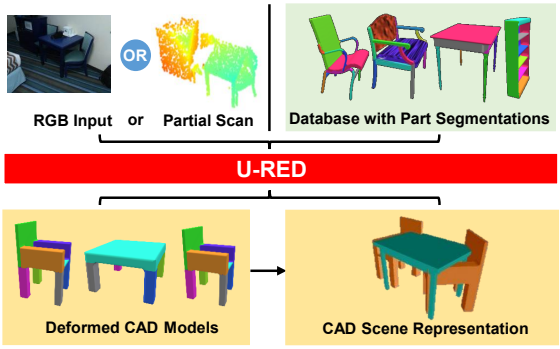

Figure 1. Given a segmented RGB image with estimated depth or a partial noisy object scan, U-RED utilizes an unsupervised joint **R**&**D** network to retrieve the most suitable CAD model from the database and deform it to tightly fit the target object. After aligning all deformed shapes to the target scene via predicted poses [18], a compact CAD scene representation is generated.

However, these methods suffer from two challenges, making them vulnerable to noise and partial observations. First, a partial shape may correspond to multiple full shapes. For example, if only a plane is observed, it may be the back

*Authors with equal contributions.
Codes: https://github.com/ZhangCYG/U-RED

or the seat of a chair. Without additional prior information, the correspondence is totally ambiguous. Directly applying supervision using a single ground truth may result in erroneous or undesired results. Therefore, the retrieval network should allow a *one-to-many* (**OTM**) retrieval. Second, due to challenging illumination conditions and inherent sensor limitations, noisy observations are common in real-world scenarios. Thereby, a learned noise-robust metric to measure similarity among shapes is essential for retrieving the most similar source shapes to the observed shape.

To handle these challenges, we propose U-RED, a novel **U**nsupervised joint 3D shape **RE**trieval and **D**eformation framework that is capable of effectively handling noisy, partial, and unseen object observations. We develop U-RED upon large-scale synthetic data that provides plenty of high-quality CAD models. We simulate real-world occlusions, sensor noise, and scan errors to generate partial point clouds of each shape as our network's input for training, then directly apply our method to challenging real-world scenes without fine tuning, since collecting 3D annotations in real scenes is laborious and requires considerable expertise.

To enable *one-to-many* retrieval, we propose to encapsulate possible full shapes of the target partial observation in the surface of a high-dimensional unit sphere. Specifically, during joint training, a supplementary branch for processing the full shape is incorporated to extract the normalized global feature, which corresponds to a point on the surface of the sphere. The full-shape feature is concatenated with the target partial-shape feature as an indicator for individual retrieval. In this manner, the retrieval network learns to interpolate different full shapes on the sphere. During inference, we sample uniformly on the sphere surface to yield multiple retrievals and collect unique ones as the final results. Moreover, cross-branch geometric consistencies can be established based on the joint learning scheme to help the partial branch learn structure-aware features and improve robustness against noise.

For similarity metrics, existing methods like [47] directly estimate a single probabilistic score based on Chamfer Distance. However, this score depends heavily on the training set and is vulnerable to noise due to the unstable nearest neighbor search in Chamfer Distance [16], limiting the generalization ability of these methods, especially in real-world scenes. We take a step further by designing a novel point-wise residual-guided metric. For each point inside the target partial observations, we predict a residual vector describing the discrepancy between the coordinates of its own and its nearest neighbor in the source shape. By aggregating all residual vectors and removing outliers, we calculate an average norm of the remaining vectors as the final metric. We demonstrate that our residual-guided metric is robust to noise and can be directly applied to real-world scenes while trained only with synthetic data.

Our main contribution are summarized as follows:

- U-RED, a novel unsupervised approach capable of conducting joint 3D shape **R**&**D** for noisy, partially-observed and unseen object observations in the real world, yielding state-of-the-art performance on public synthetic PartNet [34], ComplementMe [44] and real Scan2CAD [2] datasets.
- A novel **OTM** module that leverages supplementary full-shape cues to enable *one-to-many* retrieval and enforce geometric consistencies.
- A novel **Residual-Guided Retrieval** technique that is robust to noisy observations in real-world scenes.

2. Related Work

3D Shape Generation and Representation. Many deep learning based methods have been proposed to formulate compact representations for 3D shapes in latent space. [7, 24, 31, 32, 37, 39, 51] try to construct an implicit function by neural networks. [1, 33, 43, 52] adopt generative models to produce point clouds with high quality. Some prior works [15, 28] also present techniques of factorized representations of 3D shapes, where structural variations are modeled separately within every geometric part. Another common line of works for 3D shape representation learning [10, 26, 38, 50, 53] utilize encoder-decoder networks to generate high-dimensional latent codes which contain geometric and semantic information. A simple solution of shape retrieval task [18, 29, 47] is to directly compare the similarity of source and target shapes in latent space generated by encoder-decoder networks. However, such method is extremely sensitive to the quality of the target shapes and hard to handle partial and noisy point clouds.

CAD Model Retrieval. Retrieving a high-quality CAD model which matches tightly with a real object scan has been an important problem for 3D scene understanding. Prior static retrieval works consider the CAD-scan matching task as measuring similarity in the descriptor space [4, 40] or the latent embedding space as encoded by deep neural networks [3, 8, 18, 29]. In contrast, other approaches [35] model this task as an explicit classification problem. Since database shapes could possess the best fitting details after undergoing a deformation step, [46] proposes to extract deformation-aware embeddings, and [22] designs a novel optimization objective. However, these retrieval methods ignore the inherent connection of retrieval and deformation, leading to accumulated error and inferior performance.

3D Shape Deformation. Traditional methods proposed in the computer graphics community [14, 20, 41] directly optimize the deformed shapes to fit the input targets. However, these approaches commonly struggle to deal with real noisy and partial scans. Neural network based techniques instead try to learn deformation priors from a collection of shapes,

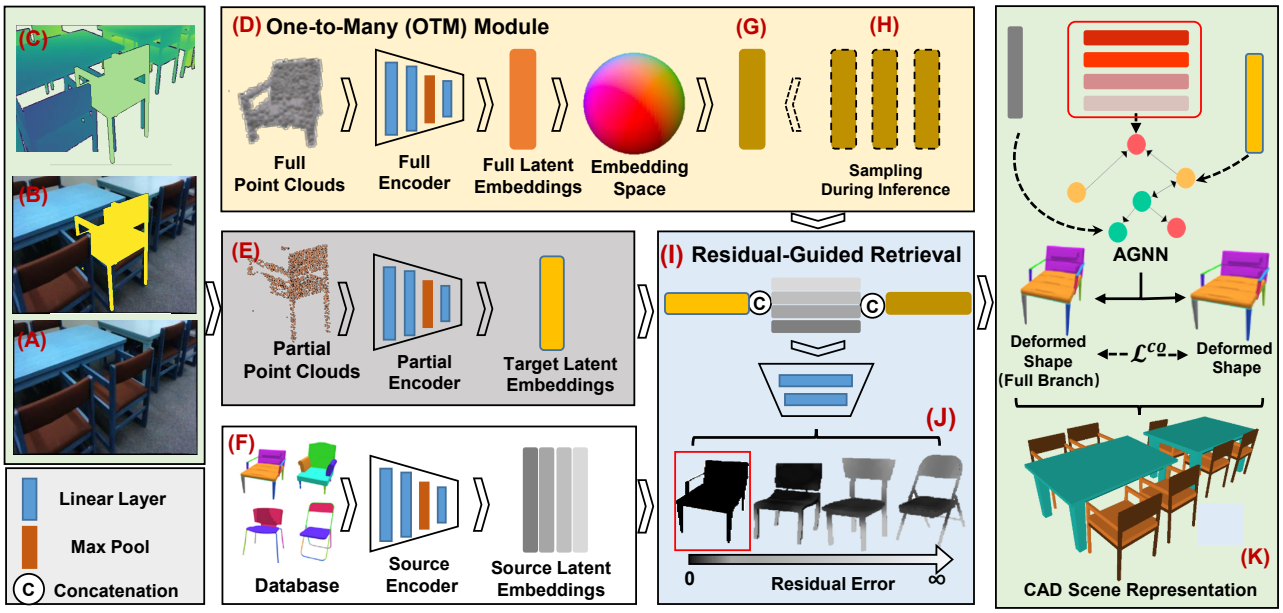

Figure 2. **Overview of U-RED.** Given an RGB image (A) that captures the target scene, we leverage an off-the-shelf object detector [19] to detect target objects (B) and a depth estimator [18] to predict the depth (C). (C) can also be directly attained via scanning. We then utilize an arbitrary pose estimator [13, 18, 59, 60] to roughly calibrate the object point cloud. Thereby partial point cloud is obtained and fed into our partial encoder (E) to extract the target feature. Subsequent retrieval module (I) takes in the target feature, source embeddings from (F) and normalized full shape feature (G), and outputs the residual field R (J) for each source shape. Note that in training, (G) is obtained via extracting latent embedding from the supplementary full shape branch, while during inference, (G) represents random samplings on the surface of a unit sphere. We choose the source shape with minimum *mean(R)* or *max(R)* as the best-fit model. Note that each sampling (G) yields a retrieval, we posit the source shape that is selected the most times as the final result. Then the retrieved source shape feature, together with the target feature and part features (in red), are optimized in AGNN [45]. The optimized part features are finally passed to an MLP to predict bounding boxes of each part. We align all the objects to generate a compact CAD-model-based scene representation.

representing deformation as volumetric warps [23, 27], cage deformations [54], vertex-based offsets [49] or flows [25]. These deformation techniques usually require constraints on grid control points, cage meshes, or number of vertices, which make them less suitable for databases with geometrically heterogeneous parts. Moreover, these assumptions are often hard to satisfy in real scans under noisy and heavily occluded settings. Recently, [47] proposes a novel training strategy with a combined loss to jointly optimize **R**&**D** at the same time. Although gaining decent performance on a synthetic dataset, its generalization ability to the partially-observed noisy scans in the real world is shown to be limited. We propose an unsupervised collaborative training technique and more tightly-coupled design for retrieval and deformation modules, yielding 3D shapes with higher quality and more precise details even when handling partial and noisy point clouds.

3. Method

3.1. Overview

Given an RGB image I or a scan $\mathcal{S}$ that captures the target scene, our method aims to detect, retrieve and deform all of its objects $O = \{O_1, ...O_N\}$, where N denotes the number of objects, and generate a clean and compact mesh-based scene representation (Fig. 2).

Feature Extraction. We stack three parallel feature encoders to extract $\{\mathcal{F}^p \in \mathbb{R}^{M\times L}, \mathcal{G}^p \in \mathbb{R}^L\}$ for the target partial point cloud $\mathcal{T}^p \in \mathbb{R}^{M\times 3}$, $\{\mathcal{F}^f \in \mathbb{R}^{M\times L}, \mathcal{G}^f \in \mathbb{R}^L\}$ for the corresponding full shape $\mathcal{T}^f \in \mathbb{R}^{M\times 3}$ (Fig. 2 (D), only in training) and $\{\mathcal{F}^d \in \mathbb{R}^{M\times L_d}, \mathcal{G}^d \in \mathbb{R}^{L_d}\}$ for sources shapes $\mathcal{O}^c$ in the database. Here $\mathcal{F}^*$ denotes the point-wise feature and $\mathcal{G}^*$ encapsulates global information. M is the number of points and $\{L, L_d\}$ are feature dimensions. Part features $\{\mathcal{P}_i^f \in \mathbb{R}^L, i = 1, 2, ..., N_p\}$ of $\mathcal{O}^c$ are computed by mean pooling $\{\mathcal{F}^d\}$ with the given part segments. N_p denotes the number of parts.

Retrieval. (Sec. 3.2) The input is the concatenation of partial features $\{\mathcal{F}^p, \mathcal{G}^p\}$, source shape features $\mathcal{G}^d$ and the normalized full shape indicator $\hat{\mathcal{G}}^f$ from $\mathcal{G}^f$ in training or a sampling $\mathcal{G}^s$ on the surface of the unit sphere Ω in inference (Fig. 2 (G, H)).

Deformation. (Sec. 3.3) Our deformation network consists of an AGNN for part-aware message propagation and an MLP for regressing the bounding box of each part. AGNN takes $\{\mathcal{P}^f, \mathcal{G}^d\}$ of the retrieved shape, as well as the target

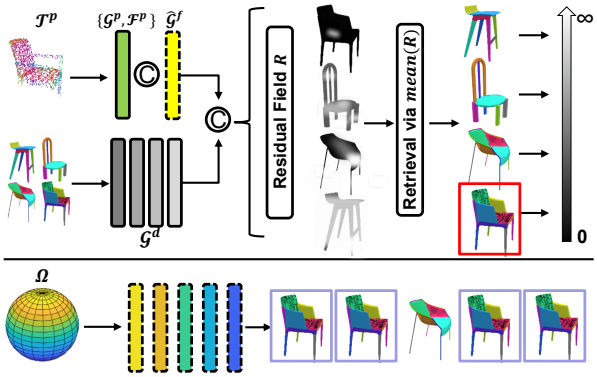

Figure 3. **Partial point cloud retrieval.** Given target partial features $\{\mathcal{G}^p, \mathcal{F}^p\}$ and candidate source shape features $\mathcal{G}^d$, we use the normalized full shape feature $\hat{\mathcal{G}}^f$ as the indicator during training, while in inference we randomly sample a feature vector from the unit sphere Ω. We estimate point-wise residual field R for each source shape, and select the object with minimum *mean*(R) (chair in the red square). Note that each sampled vector yields a retrieval result, as shown in the bottom block, we choose the object that is selected the most times in sampling as the final retrieval result (chair in the purple square).

feature $\mathcal{G}^p$ as input and outputs updated part features $\mathcal{P}'^f$, which are then fed into the MLP to predict per-part bounding boxes. The input source shape $\mathcal{O}_c$ is finally deformed to $\tilde{\mathcal{O}}^c$ that tightly matches the target $\mathcal{T}^p$ (Fig. 2 (K)).

3.2. Retrieval

It is essential for the retrieval network to learn a latent metric to measure the similarity between the target $\mathcal{T}^p$ and source shapes $\mathcal{O}_c$. Compared with previous methods leveraging full shape that encapsulates complete structures of the target , partial shape retrieval puts forward two problems.

The first problem is that one partial shape may correspond to multiple full shapes, yielding multiple feasible retrievals. Thus the retrieval network must learn a *one-to-many* relationship that enables retrieval of all possible source shapes of the target. For full shape retrieval, this can be conveniently implemented by simply employing the Chamfer Distance to supervise shape deformation and retrieval [46, 47]. However, for partial point clouds, directly applying the Chamfer Distance may lead to inaccurate nearest neighbor search and thus result in erroneous supervision. Exemplary, if only one stick is observed, it can originate from the back of a chair, yet, it can also represent a chair leg. Thereby different interpretations of the observation can lead to very different retrieval results. A straightforward idea is to harness generative point cloud completion methods [21, 55, 58] to attain full point clouds. However, this may significantly increase the inference time and limit real-world applications.

The second problem is that observed point clouds typically contain noticeable noise due to inherent sensor limitations or segmentation errors, which raises a requirement for certain robustness of the retrieval metric. Existing methods [18, 46, 47] first extract global features for the target and source shapes respectively, and then calculate retrieval probability by incorporating certain distance metrics between the features. Such strategies depend heavily on the training datasets and are vulnerable to real-world noise.

In this module, we introduce two novel techniques to solve the aforementioned problems.

To facilitate the *one-to-many* retrieval, we design a novel **OTM** module that learns to project the space of all potential full shapes $\mathcal{T}^f$ of the target partial shape $\mathcal{T}^p$ onto the surface Ω of a high-dimensional unit sphere. Specifically, during training, a supplementary branch for processing $\mathcal{T}^f$ is incorporated to extract full shape features $\{\mathcal{F}^f, \mathcal{G}^f\}$. $\mathcal{G}^f$ is normalized to be $\hat{\mathcal{G}}^f$, which corresponds to a point on Ω. Given the partial shape features $\{\mathcal{F}^p, \mathcal{G}^p\}$, we concatenate them with $\hat{\mathcal{G}}^f$ as the input of the retrieval network. $\hat{\mathcal{G}}^f$ serves as an indicator enabling *one-to-one* retrieval. The network learns to interpolate on Ω so that each point on Ω implies a possible full shape. During inference, we uniformly sample surface points on Ω as the indicator and each sampling will yield a retrieval result. We collect the unique retrievals as all feasible results and the source shape that is selected the most times is considered as the best-fit retrieval.

To robustly handle the noisy observations, we design a novel point-wise **Residual-Guided** similarity metric for **Retrieval**, as shown in Fig. 3. Our retrieval network predicts the residual field $R = \{R_i \in \mathbb{R}^3, i = 1, ..., M\}$ where for each point P_i in $\mathcal{T}^p$, its corresponding residual vector R_i describes the displacement vector from P_i to its nearest neighbor O_i in the deformed source shape $\tilde{\mathcal{O}}^c$. We adopt the $\mathcal{L}_2$ loss to supervise the training of R with

$$\mathcal{L}^{re} = \frac{1}{M}\sum_{i=1}^{M} \|P_i + R_i - Q_i\|^2, \tag{1}$$

where Q_i is obtained using nearest neighbor search. Leveraging R, we can derive $mean(R)$ for retrieval. To be robust to noise, we sort R by norm and remove 10% points with large residual norms, and then compute the average norm $mean(R)$ of the remaining points as the final metric.

3.3. Graph Attention Based Deformation

Our deformation network consists of an AGNN [45] for part-aware message aggregation and a regressor to output the bounding box of each part. The AGNN takes in 3 types of nodes. The first is global features $\mathcal{G}^d$ from the retrieved source shape $\mathcal{O}_c$. The second is global features $\mathcal{G}^p$ from the target object $\mathcal{T}^p$. The last is part features $\{\mathcal{P}_i^f, i = 1, 2, ..., N\}$ of $\mathcal{O}^c$. We stack two interleaving self-attention [48] and cross-attention modules. In self attention,

different parts exchange information, while in cross attention, global nodes propagate global structure information to guide part nodes. The overall update process is defined as,

$$\mathcal{F}' = \mathcal{F} + MHA(\mathcal{Q}, \mathcal{K}, \mathcal{V}) \tag{2}$$

where MHA refers to multi-head attention mechanism [48]. In self-attention modules, $\mathcal{Q}, \mathcal{K}, \mathcal{V} = \mathcal{F} = \mathcal{P}^f$, while in cross-attention modules, $\mathcal{Q} = \mathcal{F} = \mathcal{P}^f$, $\mathcal{K}, \mathcal{V} = \{\mathcal{G}^p, \mathcal{G}^d\}$, the concatenation of global features.

Finally, $\mathcal{P}^f$ is fed into the regressor to predict center displacement $\mathcal{C}_d$ and axis-aligned scaling parameters $\{s_w, s_h, s_l\}$. The final bounding box of each part is recovered as $\mathcal{C} = \mathcal{C}_d + \mathcal{C}_0$, $\{\mathcal{H}, \mathcal{W}, \mathcal{L}\} = \{s_w\mathcal{W}_0, s_h\mathcal{H}_0, s_l\mathcal{L}_0\}$, where $\{\mathcal{C}_0, \mathcal{W}_0, \mathcal{H}_0, \mathcal{L}_0\}$ are initial bounding box center, width, height and length respectively.

3.4. Joint Training for R&D

We utilize the synthetic dataset, PartNet [34] to generate the training data. For each available full CAD shape, we simulate real-world partial observations by random cropping and noise addition. Please refer to the Supplementary Materials for details about data generation. After training on the simulated data, we directly apply our U-RED on real-world scenes without finetuning.

Cross-Branch Consistencies. As introduced in Sec. 3.2 and Fig. 2 (D), we incorporate an additional branch to process full shapes to enable *one-to-many* retrieval. Besides this role, the full-shape branch can also be utilized to guide the partial-shape branch to understand the inherent geometric cues of the partial observations (Fig. 2 (E)). Thereby our joint learning strategy establishes explicit geometric consistencies between the two branches. For the deformation network, the retrieved source shape $\mathcal{O}^c$ is deformed to $\tilde{\mathcal{O}}_p^c$ by the partial-shape branch and $\tilde{\mathcal{O}}_f^c$ by the full shape branch. $\mathcal{O}_p^c$ should be consistent with $\mathcal{O}_f^c$. Meanwhile, given deformed source $\tilde{\mathcal{O}}_p^c$ from the partial-shape branch and $\tilde{\mathcal{O}}_f^c$ from the full-shape branch, our retrieval network yields R_p and R_f. By enforcing the consistencies between two branches, the partial-shape branch is guided to produce similar results as the full-shape branch and forced to exploit geometric characteristics of the full shape from the partial input. For each point P_i in the partial shape, its residual $\{R_{p_i} \in R_p\}$ should be consistent with the residual $\{R_{q_i} \in R_f\}$ of its corresponding point Q_i in the full shape. We define $R'_f = \{R_{q_i}\}$, and thus our cross-branch consistency loss is defined as,

$$\mathcal{L}^{co} = \mathcal{L}_1^{co} + \mathcal{L}_2^{co} = \|\mathcal{O}_p^c - \mathcal{O}_f^c\|^2 + \|\mathcal{R}_p - \mathcal{R'}_f\|^2 \tag{3}$$

where we use the Euclidean Distance between point sets as the loss function.

Overall Objective. Aggregating all losses used in our **R**&**D** framework, the final objective for our unsupervised training can be summarized as,

$$\mathcal{L} = \lambda_0\mathcal{L}^b + \lambda_1\mathcal{L}^{re} + \lambda_2\mathcal{L}^{co} \tag{4}$$

where $\{\lambda_0, \lambda_1, \lambda_2\}$ are weighting parameters. $\mathcal{L}^b$ contains several basic losses, including Chamfer Distance loss $\mathcal{L}^{cd}$ and reconstruction loss $\mathcal{L}^r$. We provide more details in the Supplementary Material.

4. Experiments

We mainly conduct experiments on partial input in this section. For results of full shape input, please refer to the Supplementary Material.

4.1. Experimental Setup

Dataset Preparation. We evaluate U-RED on 3 public datasets. In particular, two synthetic datasets PartNet [34], ComplementMe [44] and one real-world dataset Scan2CAD [2]. PartNet is associated with ShapeNet [6] and provides fine-grained part segments for 3 furniture categories: chairs (6531), tables (7939) and cabinets (1278). ComplementMe contains automatically-segmented shapes of two categories, *i.e.* chairs and tables. To further demonstrate the generalization ability and practical utility of U-RED, we also adopt a real-world dataset Scan2CAD which is derived from ScanNet [9], to directly evaluate our trained model on synthetic data without finetuning. We follow [47] and divide PartNet and ComplementMe into database, training and testing splits. Synthetic partial inputs for training are generated by simulating occlusions and sensor errors on original full CAD shapes in PartNet and ComplementMe. For Scan2CAD, we adopt the test split of ROCA [18] and utilize data of three categories: chairs, tables and cabinets.

Database Construction. We utilize PartNet and ComplementMe to establish the database. Following [47], we randomly selected 10% of the data from the finest level of PartNet hierarchy and additionally select 200 chairs and 200 tables from ComplementMe. Throughout all experiments, we always rely on the same database.

Implementation Details. We represent all shapes by uniformly sampling $M = 1024$ points. We use the AdamW optimizer with initial learning rate $1e-3$ and train U-RED for 200 epochs. For loss weights, we set $\{\lambda_0, \lambda_1, \lambda_2\} = \{3.0, 0.3, 1.0\}$ unless specified. For all extracted pointwise, global or part features, we set the feature dimension as $L = L_d = 256$. The supplementary full-shape branch share the hyperparameter setting with the partial branch. For retrieval implementation, for fair comparison to Top-K soft retrieval adopted by [47], we also take Top-K source candidates with most selected times as described in Sec. 3.2. In all experiments, we sample 1000 times on the unit sphere to generate 1000 retrieval results and use top-10 candidates.

Method	Chair	Table	Cabinet	Average
Uy et al. [47]	3.36	6.65	7.26	4.90
ROCA* [18]	4.24	14.97	15.92	9.15
ROCA* [18]+De	6.99	8.10	13.08	8.10
Ours	**2.89**	**3.16**	**5.95**	**3.35**

Table 1. Joint **R**&**D** results on real Scan2CAD [2] dataset. Our U-RED achieves 31.6% leap forward over competitors of under the instance-average Chamfer Distance.

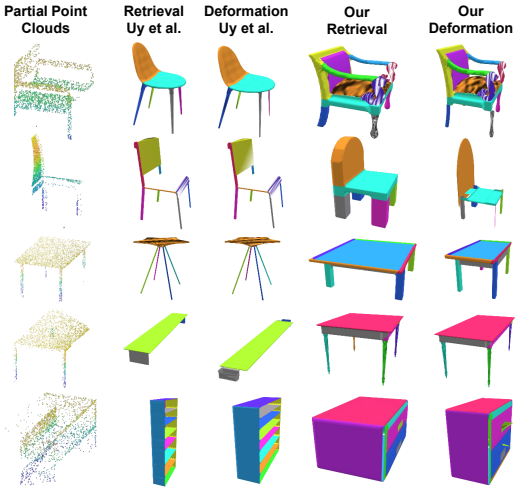

Figure 4. Qualitative comparison on **R**&**D** results with Uy *et al.* [47] on real-world Scan2CAD [2].

Evaluation Metrics. We report the average Chamfer Distance on the magnitude of 10^{-2} of each category respectively and then estimate the instance-average Chamfer Distance on the whole dataset. The Chamfer Distance is calculated between the deformed source shapes and the ground truth shapes.

Baseline Methods. We train and evaluate all baseline methods [18, 47] and U-RED with the same data split, with our genarated partial shapes. All baseline methods and their variants are trained with the parameters reported in the original papers until convergence. We do not use the inner deformation optimization (IDO) step of Uy *et al.* [47] for 2 reasons: First, IDO significantly increases the training time by almost 10 times. Second, IDO works only on the deformation module as a refinement strategy, and can also be incorporated into U-RED. Therefore, we focus on comparing the effectiveness of the basic networks and methods, rather than the results after refinement.

4.2. Real-world Scenes

We first test our proposed approach on the real-world scenes. We train U-RED on the synthetic PartNet and test it on the real-world Scan2CAD dataset without finetuning or retraining. We present the respective results in Tab. 1, which demonstrates that U-RED yields the most precise deformed CAD models. U-RED shows strong generalization ability, which attributes to our elaborated module design and effective unsupervised collaborative learning.

In particular, when comparing with [47], we report superior results on the real-world scenes with decreased Chamfer Distance by 14.0%, 52.5% and 18.0% for three categories respectively. Moreover, a qualitative comparison is shown in Fig. 4, and Fig. 5. It can be easily seen that our approach yields more accurate retrieval results as well as more precise deformations with respect to the actual real-world objects. Note that for a fair comparison, we train [47] with our partial synthetic objects in order to improve its results on partially observed point clouds. The noticeable noise in real-world scenes easily deteriorates its performance.

Further, we re-implement ROCA [18] by preserving only the retrieval head, notated as ROCA*. For a fair comparison, we also present results for our improved ROCA*+De, for which we concatenate our deformation head to ROCA*. We train ROCA* and ROCA*+De on the same synthetic PartNet dataset and test them on Scan2CAD the same as U-RED. Compared with ROCA*, the Chamfer Distance yielded by U-RED decreases by 63.3% on average. When compared to ROCA*+De, we still have 58.6% performance gain (See Tab. 1). Although the performance of ROCA* improves somewhat when tested on the synthetic dataset, as shown in Tab. 2, it is incapable of conducting proper domain adaptation. Despite adding our deformation head in ROCA*+De, the weak retrieval results significantly increase the difficulty of the deformation module, resulting in only a minor improvement, from 9.15 to 8.10. In contrast, our residual-guided retrieval technique possesses superior noise resistance and enforces effective learning of geometric cues due to the collaborative training procedure. Thus, it mitigates the negative effects of noisy input and provides robust results for real-world scans.

In Fig. 8, we demonstrate the qualitative comparisons of Ours *vs* Uy *et al.* on Scan2CAD with only RGB input. Specifically, given an RGB image as input, we detect the target objects with Mask-RCNN [19] and predict depth with ROCA [18]. The recovered point cloud is then transformed via the estimated pose from ROCA and used as input to the methods.

4.3. Synthetic Scenes

The objective of U-RED is to enhance the generalization ability with unsupervised training procedures to handle noisy and partial input in real-world scenes. However, we still conduct several experiments on synthetic datasets for a fair comparison with existing methods [18, 47]. The experiment results (See Tab. 2) show that our U-RED also outperforms the state-of-the-art competitors by a large margin

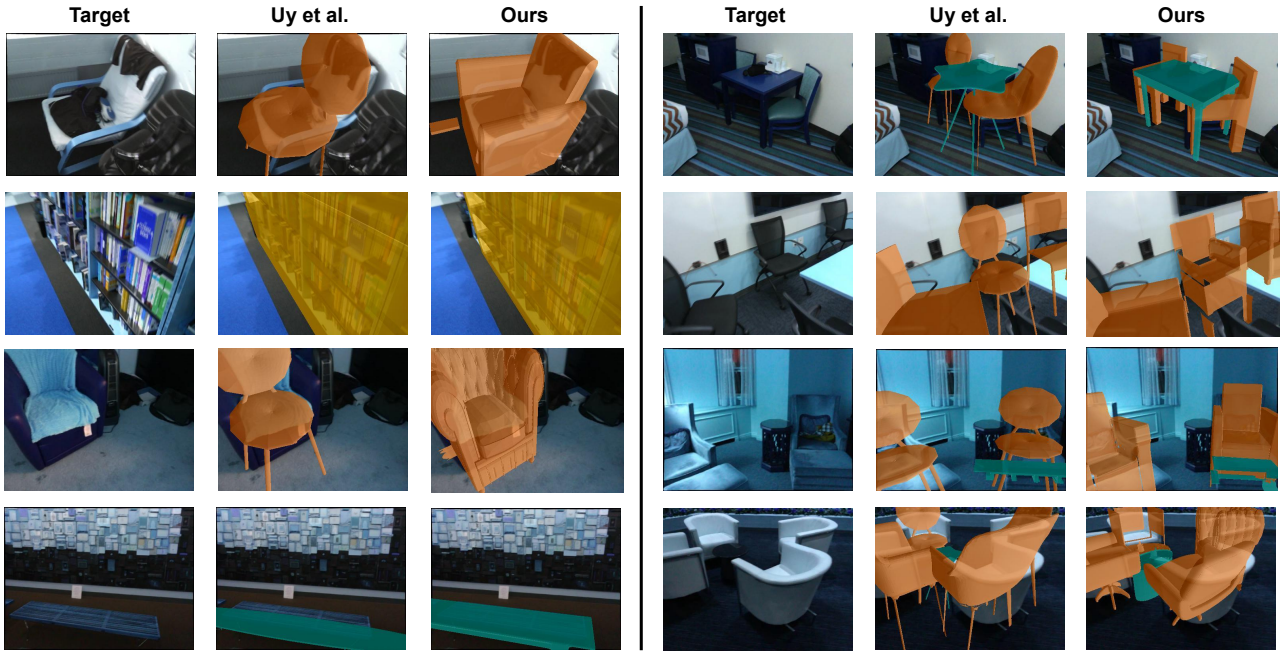

Figure 5. Qualitative results on Scan2CAD [2] dataset. U-RED consistently outperforms state-of-the-art Uy *et al.* [47].

PartNet [34]				
Method	Chair	Table	Cabinet	Average
Uy et al. [47]	2.02	2.32	2.63	2.22
ROCA* [18]	2.50	2.72	3.86	2.72
ROCA* [18]+De	3.80	3.87	2.82	3.76
Ours	**0.95**	**1.33**	**1.30**	**1.17**
ComplementMe [44]				
Method	Chair	Table	-	Average
Uy et al. [47]	2.08	2.66	-	2.40
Ours	**1.68**	**2.26**	-	**2.00**

Table 2. Joint **R**&**D** results on synthetic datasets. Our U-RED outperforms all competitors in synthetic scenes with both manually and automatically segmentation.

when dealing with ambiguous and partially observed point clouds in synthetic scenes.

On PartNet, U-RED surpasses [47] by 47.3% on average of the three categories, and exceeds ROCA* and ROCA*+De by an even larger margin. On ComplementMe, U-RED still surpasses [47] by 19.2% and 15.0% for chair and table categories respectively. Fig. 6 exhibits the qualitative visualization comparison between our U-RED and [47]. U-RED demonstrates more accurate retrieval results (the chair in the 1^{st} row) and stronger deformation capability (the table in the 5^{th} row). Fig. 7 demonstrates the results with only RGB input, with the same setting as in Fig. 8.

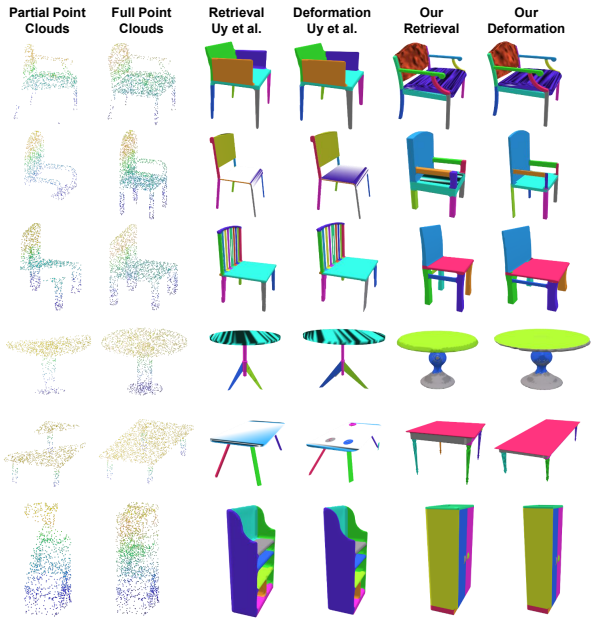

Figure 6. Visualization on PartNet [34]. Qualitative comparison with Uy *et al.* [47] demonstrates that our U-RED performs more robustly, gains more accurate retrieval and more precise deformation results when facing partially observed point clouds.

4.4. Ablation Studies

We conduct several ablation studies on deformation ability (Tab. 3), retrieval ability (Tab. 4), the effectiveness of

Scan2CAD [2]				
Method	Chair	Table	Cabinet	Average
Uy et al. [47]	2.83	2.47	**4.56**	2.92
Ours	**2.05**	**1.43**	5.27	**2.24**
PartNet [34]				
Method	Chair	Table	Cabinet	Average
Uy et al. [47]	1.43	1.48	1.79	1.48
Ours	**0.70**	**0.69**	**0.74**	**0.70**

Table 3. Deformation results with oracle retrieval. Results on two datasets show our U-RED gains stronger deformation ability due to our graph attention based deformation module.

Scan2CAD [2]				
Method	Chair	Table	Cabinet	Average
Pro-Re [47]	7.39	6.00	15.24	5.95
Ours w/o **OTM**	3.19	6.62	7.02	4.71
Ours	**2.89**	**3.16**	**5.95**	**3.35**
PartNet [34]				
Method	Chair	Table	Cabinet	Average
Pro-Re [47]	2.84	5.17	4.66	4.15
Ours w/o **OTM**	1.17	2.16	1.85	1.76
Ours	**0.95**	**1.33**	**1.30**	**1.17**

Table 4. Retrieval ablations. Ours w/o **OTM** refers to our method without the **OTM** module. Results illustrate that our residual-guided retrieval gains more precise retrieval for the targets than the probabilistic retrieval adopted by [47].

Scan2CAD [2]				
Method	Chair	Table	Cabinet	Average
w/o $\mathcal{L}^{co}$	3.57	5.56	6.54	4.58
w/o $\mathcal{L}^{r}$	4.47	5.03	9.11	5.22
w/o $\mathcal{L}^{re}$	2.95	5.32	19.6	5.75
Ours	**2.89**	**3.16**	**5.95**	**3.35**
PartNet [34]				
Method	Chair	Table	Cabinet	Average
w/o $\mathcal{L}^{co}$	1.01	1.41	2.25	1.31
w/o $\mathcal{L}^{r}$	1.37	1.63	1.51	1.51
w/o $\mathcal{L}^{re}$	2.02	1.77	1.54	1.86
Ours	**0.95**	**1.33**	**1.30**	**1.17**

Table 5. Unsupervised training ablations. The results prove that each single loss term in our unsupervised training procedure contributes to better performance.

unsupervised joint training techniques (Tab. 5) and robustness to input occlusion proportion (Tab. 6). These ablation studies demonstrate the effectiveness of U-RED in handling noisy partial points in both real-world and synthetic scenes.

Deformation Ability. In this experiment, we research on the deformation ability alone using oracle retrieval, i.e. the deformation network deforms every source shape in the database to match a target shape and reports the minimum Chamfer Distances between the target shape and all deformed shapes. We compare the deformation ability of our graph attention-based deformation module with [47]. Tab. 3 shows that when only considering deformation ability, U-RED surpasses [47] by 23.3% on average on real-world Scan2CAD and 52.7% on synthetic PartNet.

Occlusion	Chair	Table	Cabinet	Average
0%	0.77	1.33	1.18	1.08
25%	0.86	1.33	1.23	1.12
50%	0.95	1.33	1.30	1.17
75%	1.09	2.21	1.65	1.69

Table 6. Occlusion ablations on PartNet [34]. Results illustrate that our U-RED performs stably even under heavy occlusion.

Retrieval Ability. We aim to verify the effectiveness of our proposed *one-to-many* retrieval and residual-guided retrieval metric. As in Tab. 4, Ours w/o **OTM** refers to persevering the residual-guided metric but turn off the **OTM** module. Thereby the normalized full shape feature $\hat{\mathcal{G}}^f$ is not fed into the retrieval network. Moreover, we need a baseline that doesn't leverage both the two techniques. For this purpose, we implemented a network with the probabilistic soft retrieval strategy from [47] and leverages other modules the same as U-RED, notated as Probabilistic Retrieval (Pro-Re). Tab. 4 verifies the effectiveness of our proposed residual-guided retrieval module. Compare Pro-Re and Ours w/o **OTM**, it can be easily concluded that the residual-guided metric improves the retrieval accuracy remarkably by 20.8% on Scan2CAD and 56.8% on PartNet. When comparing ours w/o **OTM** and Ours, it's clear that the **OTM** retrieval strategy contributes 28.9% and 33.5% improvements on Scan2CAD and PartNet respectively. We provide several qualitative retrieval results to further demonstrate the effectiveness of **OTM** in the Supplementary Material.

Loss terms. We ablate the effectiveness of each proposed loss term on both real and synthetic scenes and summarize the experiment results in Tab. 5. As illustrated, we proved that each single loss term contributes to better performance. And finally, the network achieve the best performance by applying of all the loss terms.

Input Occlusion Proportion. Tab. 6 examines the robustness of our method when tackling input partially-observed point clouds with different occlusion proportions. The occlusion proportions are controlled by our simulation technique adopted to generate the partial point clouds input. Our U-RED performs stably even with an occlusion ratio of 50%. We observe an accuracy drop for the first time when the occlusion ratio increase to 75%. The result verifies the robustness of our approach.

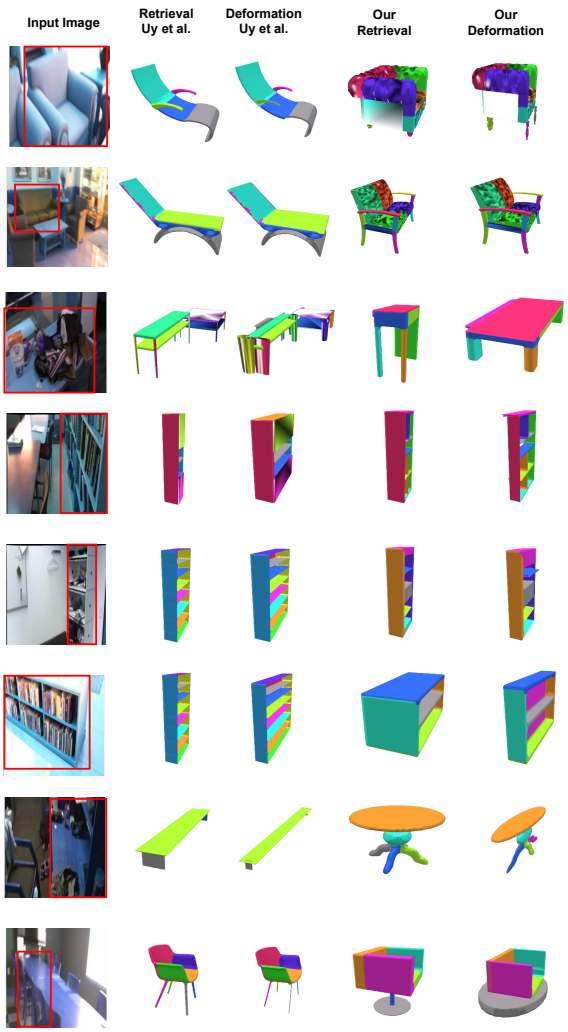

Figure 7. Qualitative results on Scan2CAD [2] dataset with only RGB input. We leverage an off-the-shelf object detector [19] to detect target objects and a depth estimator [18] to predict the depth. Then we use our U-RED for 3D shape retrieval and deformation. U-RED consistently outperforms state-of-the-art Uy *et al.* [47] considering visual effects.

We also provide experiments analyzing the effects of noise level, and different retrieval metrics: min *mean(R) vs.* min *max(R)* in the Supplementary Material.

5. Conclusion

In this paper, we present U-RED, a novel unsupervised framework that takes a partial object observation as input, aiming to retrieve the most geometrically similar shape from a pre-established database and deform the retrieved shape to match the target. To solve the one-to-many retrieval problem and enhance the robustness to handle noisy partial points, we design a collaborative unsupervised learning technique with the aid of a supplementary full-shape branch, enforcing geometric consistencies. Further, we propose a residual-guided retrieval module that is robust to noisy observations and a graph attention based deformation module for gaining more precise deformed shapes. The training technique and all proposed modules are demonstrated to be effective through our exhaustive experiments in both real-world and synthetic scenes. In the future, we plan to apply our retrieval technique in a wide range of real-world applications, *e.g.* model-based robotic grasping.

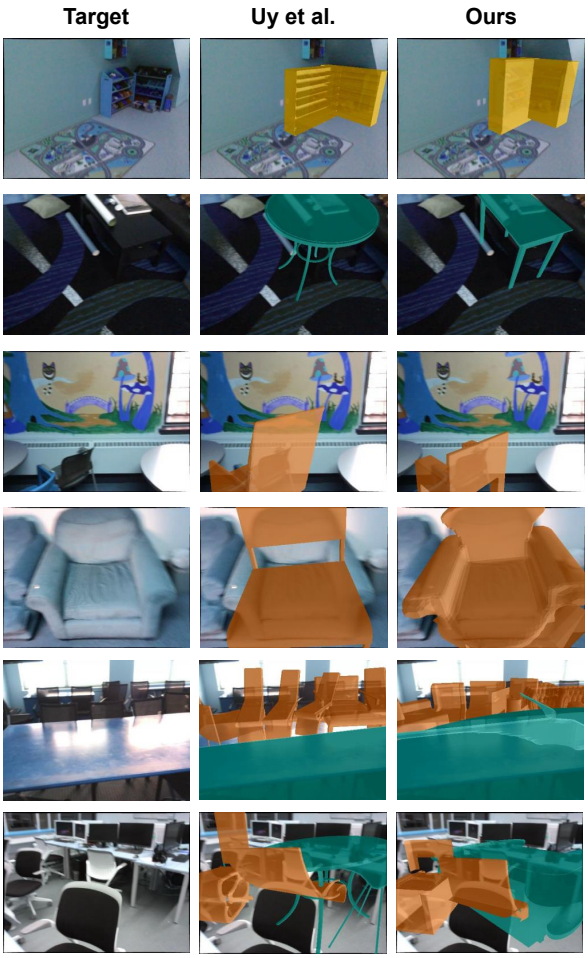

Figure 8. Qualitative results on Scan2CAD [2] dataset with only RGB input. U-RED consistently outperforms state-of-the-art Uy *et al.* [47].

Acknowledgements. This work was partially funded by the German Federal Ministry for Economics and Climate Action (BMWK) under GA 13IK010F (TWIN4TRUCKS) and the EU Horizon Europe Framework Program under GA 101058236 (HumanTech). This work was also supported by the National Key R&D Program of China under Grant 2018AAA0102801, National Natural Science Foundation of China under Grant 61827804.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 2

[2] 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, 5, 6, 7, 8, 9

[3] Armen Avetisyan, Angela Dai, and Matthias Nießner. End-to-end cad model retrieval and 9dof alignment in 3d scans. In *Proceedings of the IEEE/CVF International Conference on computer vision*, pages 2551–2560, 2019. 2

[4] Frederic Bosche and Carl T Haas. Automated retrieval of 3d cad model objects in construction range images. *Automation in Construction*, 17(4):499–512, 2008. 2

[5] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1

[6] 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. 5

[7] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 2

[8] Manuel Dahnert, Angela Dai, Leonidas J Guibas, and Matthias Nießner. Joint embedding of 3d scan and cad objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8749–8758, 2019. 1, 2

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 5

[10] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5868–5877, 2017. 2

[11] Yan Di, Henrique Morimitsu, Shan Gao, and Xiangyang Ji. Monocular piecewise depth estimation in dynamic scenes by exploiting superpixel relations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4363–4372, 2019. 1

[12] Yan Di, Henrique Morimitsu, Zhiqiang Lou, and Xiangyang Ji. A unified framework for piecewise semantic reconstruction in dynamic scenes via exploiting superpixel relations. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10737–10743. IEEE, 2020. 1

[13] 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

[14] Vignesh Ganapathi-Subramanian, Olga Diamanti, Soeren Pirk, Chengcheng Tang, Matthias Niessner, and Leonidas Guibas. Parsing geometry using structure-aware shape templates. In *2018 International Conference on 3D Vision (3DV)*, pages 672–681. IEEE, 2018. 2

[15] Lin Gao, Jie Yang, Tong Wu, Yu-Jie Yuan, Hongbo Fu, Yu-Kun Lai, and Hao Zhang. Sdm-net: Deep generative network for structured deformable mesh. *ACM Transactions on Graphics (TOG)*, 38(6):1–15, 2019. 2

[16] Chris R. Giannella. Instability results for euclidean distance, nearest neighbor search on high dimensional gaussian data. *Information Processing Letters*, 169:106115, 2021. 2

[17] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9785–9795, 2019. 1

[18] 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, 2, 3, 4, 5, 6, 7, 9

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3, 6, 9

[20] Qi-Xing Huang, Bart Adams, Martin Wicke, and Leonidas J Guibas. Non-rigid registration under isometric deformations. In *Computer Graphics Forum*, volume 27, pages 1449–1457. Wiley Online Library, 2008. 2

[21] Zitian Huang, Yikuan Yu, Jiawen Xu, Feng Ni, and Xinyi Le. Pf-net: Point fractal network for 3d point cloud completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7662–7670, 2020. 4

[22] Vladislav Ishimtsev, Alexey Bokhovkin, Alexey Artemov, Savva Ignatyev, Matthias Niessner, Denis Zorin, and Evgeny Burnaev. Cad-deform: Deformable fitting of cad models to 3d scans. In *European Conference on Computer Vision*, pages 599–628. Springer, 2020. 2

[23] Dominic Jack, Jhony K Pontes, Sridha Sridharan, Clinton Fookes, Sareh Shirazi, Frederic Maire, and Anders Eriksson. Learning free-form deformations for 3d object reconstruction. In *Asian Conference on Computer Vision*, pages 317–333. Springer, 2018. 3

[24] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021. 2

[25] 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. 1, 3

[26] Jincen Jiang, Xuequan Lu, Lizhi Zhao, Richard Dazeley, and Meili Wang. Masked autoencoders in 3d point cloud representation learning. *arXiv preprint arXiv:2207.01545*, 2022. 2

[27] Andrey Kurenkov, Jingwei Ji, Animesh Garg, Viraj Mehta, JunYoung Gwak, Christopher Choy, and Silvio Savarese. Deformnet: Free-form deformation network for 3d shape reconstruction from a single image. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 858–866. IEEE, 2018. 3

[28] Jun Li, Kai Xu, Siddhartha Chaudhuri, Ersin Yumer, Hao Zhang, and Leonidas Guibas. Grass: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 2

[29] Yangyan Li, Hao Su, Charles Ruizhongtai Qi, Noa Fish, Daniel Cohen-Or, and Leonidas J Guibas. Joint embeddings of shapes and images via cnn image purification. *ACM transactions on graphics (TOG)*, 34(6):1–12, 2015. 2

[30] Jiachen Liu, Pan Ji, Nitin Bansal, Changjiang Cai, Qingan Yan, Xiaolei Huang, and Yi Xu. Planemvs: 3d plane reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8665–8675, 2022. 1

[31] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2

[32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[33] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas J Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575*, 2019. 2

[34] 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, 5, 7, 8

[35] Liangliang Nan, Ke Xie, and Andrei Sharf. A search-classify approach for cluttered indoor scene understanding. *ACM Transactions on Graphics (TOG)*, 31(6):1–10, 2012. 1, 2

[36] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 1

[37] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[38] Yuchen Rao, Yinyu Nie, and Angela Dai. Patchcomplete: Learning multi-resolution patch priors for 3d shape completion on unseen categories. *arXiv preprint arXiv:2206.04916*, 2022. 2

[39] Edoardo Remelli, Artem Lukoianov, Stephan Richter, Benoît Guillard, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Meshsdf: Differentiable iso-surface extraction. *Advances in Neural Information Processing Systems*, 33:22468–22478, 2020. 2

[40] Adriana Schulz, Ariel Shamir, Ilya Baran, David IW Levin, Pitchaya Sitthi-Amorn, and Wojciech Matusik. Retrieval on parametric shape collections. *ACM Transactions on Graphics (TOG)*, 36(1):1–14, 2017. 1, 2

[41] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116, 2007. 2

[42] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 1

[43] Yongbin Sun, Yue Wang, Ziwei Liu, Joshua Siegel, and Sanjay Sarma. Pointgrow: Autoregressively learned point cloud generation with self-attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 61–70, 2020. 2

[44] Minhyuk Sung, Hao Su, Vladimir G Kim, Siddhartha Chaudhuri, and Leonidas Guibas. Complementme: Weakly-supervised component suggestions for 3d modeling. *ACM Transactions on Graphics (TOG)*, 36(6):1–12, 2017. 2, 5, 7

[45] Kiran K Thekumparampil, Chong Wang, Sewoong Oh, and Li-Jia Li. Attention-based graph neural network for semi-supervised learning. *arXiv preprint arXiv:1803.03735*, 2018. 3, 4

[46] Mikaela Angelina Uy, Jingwei Huang, Minhyuk Sung, Tolga Birdal, and Leonidas Guibas. Deformation-aware 3d model embedding and retrieval. In *European Conference on Computer Vision*, pages 397–413. Springer, 2020. 1, 2, 4

[47] 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, 5, 6, 7, 8, 9

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 5

[49] Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 3dn: 3d deformation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1038–1046, 2019. 1, 3

[50] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2690–2698, 2019. 2

[51] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 2

[52] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4541–4550, 2019. 2

[53] 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. 2

[54] Wang Yifan, Noam Aigerman, Vladimir G Kim, Siddhartha Chaudhuri, and Olga Sorkine-Hornung. Neural cages for detail-preserving 3d deformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 75–83, 2020. 1, 3

[55] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12498–12507, 2021. 4

[56] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8833–8842, 2021. 1

[57] Chenyangguang Zhang, Zhiqiang Lou, Yan Di, Federico Tombari, and Xiangyang Ji. Sst: Real-time end-to-end monocular 3d reconstruction via sparse spatial-temporal guidance. *arXiv preprint arXiv:2212.06524*, 2022. 1

[58] Junzhe Zhang, Xinyi Chen, Zhongang Cai, Liang Pan, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, Bo Dai, and Chen Change Loy. Unsupervised 3d shape completion through gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1768–1777, 2021. 4

[59] 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

[60] 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

[61] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 1