Patch2CAD: Patchwise Embedding Learning for In-the-Wild Shape Retrieval from a Single Image

Weicheng Kuo[1], Anelia Angelova[1], Tsung-Yi Lin[1], Angela Dai[2]
[1] Google Research, Brain Team
[2] Technical University of Munich
{weicheng, anelia, tsungyi}@google.com, angela.dai@tum.de

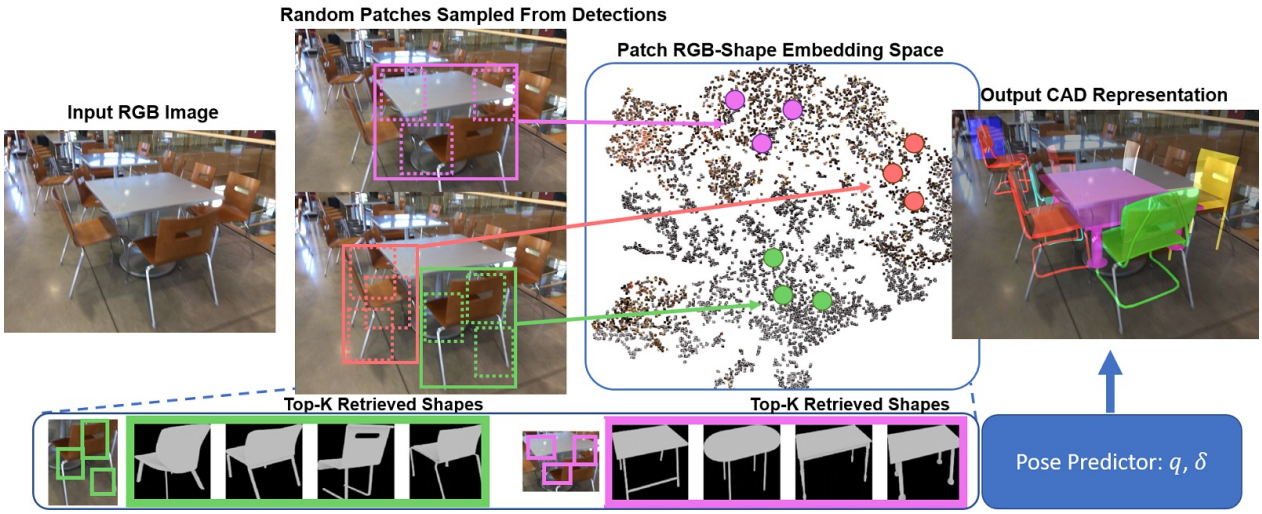

Figure 1: From an input RGB image, we learn a shared image-CAD embedding space by embedding patches of detected objects from the RGB images and patches of CAD models. By establishing patch-wise correspondence between image and CAD, we can establish object correspondence based on part similarities, enabling more effective shape retrieval for new views, as well as robust top-k CAD retrieval. Our patch-based retrieval for a similar 3D CAD representation coupled with pose prediction results in a CAD-based 3D understanding of the objects in the image.

Abstract

3D perception of object shapes from RGB image input is fundamental towards semantic scene understanding, grounding image-based perception in our spatially 3-dimensional real-world environments. To achieve a mapping between image views of objects and 3D shapes, we leverage CAD model priors from existing large-scale databases, and propose a novel approach towards constructing a joint embedding space between 2D images and 3D CAD models in a patch-wise fashion – establishing correspondences between patches of an image view of an object and patches of CAD geometry. This enables part similarity reasoning for retrieving similar CADs to a new image view without exact matches in the database. Our patch embedding provides more robust CAD retrieval for shape estimation in our end-to-end estimation of CAD model shape and pose for detected objects in a single input image. Experiments on in-the-wild, complex imagery from ScanNet show that our approach is more robust than state of the art in real-world scenarios without any exact CAD matches.

1. Introduction

Fundamental to many visual perception tasks is an understanding of the decomposition of an observed scene into its constituent objects, and the semantic meaning of these objects – including class categorization, segmentation, and structural 3D understanding. In recent years, advances in 2D object recognition and localization have achieved impressive success in image-based understanding, even from only single image input [22, 18, 46, 21]. Such recognition and perception constrained to the image domain unfortunately remains limited towards understanding 3D attributes

such as shape and structure, which are not only fundamental towards a comprehensive, human-like understanding of objects in a scene, but towards many applications such as autonomous exploration and interaction of an environment.

To address 3D perception from a single RGB image, we have recently seen several methods proposed taking a generative approach towards reconstructing the observed objects' geometry [19, 42, 14]. These methods show promising results in attaining 3D understanding of objects in complex scene imagery, but take a low-level approach towards geometric reconstruction, constructing voxel-by-voxel or vertex-by-vertex, often resulting in noisy or oversmoothed geometry, or geometry not representing a valid object instance (*e.g.*, missing a leg on a chair). In contrast, several approaches have taken a CAD-based prior for representing the 3D structure of the objects seen in an RGB or RGB-D observation, by retrieving and aligning CAD models from a database similar to the observed objects [41, 33, 2, 30]. This CAD prior enables representation of each object with a clean, complete 3D mesh known to represent valid instances of objects and able to be stored compactly for potential downstream applications. Unfortunately, such a retrieval-based approach tends to struggle with generalization, in particular when a new observed image of an object does not exactly match any CAD model in the dataset.

We observe that in these challenging scenarios, various part similarities can be leveraged to find a similar shape. Thus, we propose Patch2CAD, which constructs a joint embedding space between images and CAD models based on encoding mid-level geometric relations by establishing similarity of patches of images to patches of object geometry. These correspondences can be aggregated into CAD prediction by majority voting. This enables CAD retrieval based on the predominant part similarities, enabling improved generalizability for CAD retrieval to reconstruct the shapes of objects seen in an image.

To achieve a 3D understanding of object structure from a single RGB image, we first detect object locations in 2D, then construct our patch-based image-CAD embedding space enabling voting for retrieval of a similar CAD model, and predict the pose of that CAD in the image. Patch2CAD is trained end-to-end to comprehensively establish an effective image-CAD embedding.

Our main contribution is a patch-based learning of a joint embedding space between the two very different domains of 2D images and 3D CAD models, which establishes more robust, part-level-based correspondences (see Figure 1). We demonstrate that this patch-wise embedding enables meaningful CAD retrievals for image observations not just in the top nearest neighbor, but for top-k retrieval. As a result, we achieve more effective association of CAD shapes to images observations of objects with no exact CAD matches in a candidate database, as is typically the case for real-world scenarios. We demonstrate Patch2CAD's effective 3D shape perception on both ScanNet [11] and Pix3D [52] datasets. In particular, on the complex, in-the-wild images from ScanNet [11], Patch2CAD exhibits notable advantage to its patch-based approach, outperforming state of the art by 1.9 Mesh AP (22% relative improvement).

2. Related Work

Scene understanding is one of the fundamental problems in computer vision. A vast amount of literature on the topic has forwarded the field in understanding of 2D images: for example popular methods for object detection [18, 46, 45, 37, 36, 32, 15, 58], semantic segmentation [38, 23] and instance segmentation [21, 31]. Our approach is inspired by these 2D image understanding approaches, but instead focuses on producing a 3D representation of the objects observed in a single image, providing additional geometric, structural information about the scene.

Single-View Object Reconstruction. Recently, we have also seen remarkable progress in reconstructing the 3D shape of an object from a single RGB image. Such work has also been driven by shape representations: earlier research focused on dense volumetric grids [10, 57], while point clouds [16, 59] and hierarchical structures such as octrees [53, 47] offered more memory and computationally efficient representations. Mesh-based approaches offer an efficient surface representation along with adaptive structure, but tend to rely on strong topological assumptions, taking a deformation-based approach from a given template mesh [55, 56]; generative approaches without relying on templates tend to be limited to small numbers of vertices [12]. Implicit functions have recently seen notable success in single-object shape reconstruction, characterizing shape by prediction an occupancy or signed distance field value for a location in space [40, 44, 49]. Approaches to predict convex primitives have also been shown to produce promising results [13].

While these approaches operated on images encompassing only one object, Mesh R-CNN pioneered an approach for generating the shapes for multiple objects seen in an RGB image, which more closely represents real-world perception scenarios. Several methods have now furthered development on this task; Mask2CAD [30] proposed a CAD retrieval-based approach towards understanding the shape and pose of objects, and Nie *et al*. [42] a mesh generation approach for object reconstruction based on initial deformation from a sphere followed by edge refinement to handle local topology errors. Our approach also tackles shape reconstruction for the multiple objects seen in an RGB image, leveraging CAD retrieval and focusing on the construction of a robust image-CAD embedding space.

CAD-Based Retrieval and Alignment. An alternative to generative methods for reconstruction is to leverage CAD model priors to represent objects in a scene, and retrieve and align them to achieve a scene reconstruction composed of clean, compact mesh representations of each object. Early work in computer vision demonstrated the use of existing geometric models as priors [6, 9, 48]; the current availability of large-scale CAD model datasets (*e.g.*, ShapeNet [7], Pix3D [52]) has revitalized this approach. Various methods have been introduced for CAD model retrieval and alignment to RGB-D scans [50, 29, 33, 5, 2, 20, 27], including end-to-end learning pipelines [3, 4], as well as CAD alignment to an image assuming that shape is given [35, 17, 24].

From a single image, Aubry et al. [1] develop hand-crafted HOG-based features to match textured renderings of CAD models to images in order detect chairs; our approach learns to associate CAD patches with image patches based on a more general learning of purely geometric correspondence, enabling learning geometric structures as well as the use of geometry-only CAD databases. More recently, Izadinia and Seitz [28] and Huang *et al*. [25] apply analysis-by-synthesis approaches for CAD model alignment and scene layout estimation from a single image, leveraging a costly optimization (minutes to hours) for each input image.

Shape retrieval methods also show promising results by learning joint RGB-CAD space embedding [54, 34, 39, 30]. Li *et al*. [34] propose a method to construct a joint embedding space between RGB images and CAD models, enabling CAD model retrieval from images; the embedding space is first constructed from shape descriptors, and then image embeddings are optimized for into the shape space. Massa *et al*. [39] learns to adapt object RGB features to CAD space with a projection layer for object instance detection. Kuo *et al*. [30] jointly optimize for a shared embedding space between image views and CAD models in order to perform retrieval for multiple objects seen in an image. Such techniques can be prone to overfitting, as an object shape obtains a single, global representation, and a new image may not contain exact CAD matches but rather various part similarities. Our approach addresses a similar problem statement in learning a mapping between images and CAD models; however, to better generalize to new observations with inexact matches, leveraging a majority part similarity to more robustly retrieve CAD models for reconstruction.

3. Method

3.1. Overview

From a single input RGB image, we aim to understand the observed scene by predicting the object semantics and 3D structures, by retrieving and aligning similar CAD models to the observed image. Objects are first detected in the 2D image, represented by their 2D bounding box, class label, and 2D instance segmentation mask. We then aim to learn a shared embedding space between the image representation of the objects and CAD models, in order to retrieve a similar CAD model representing the 3D structure of a detected object. In a separate pose-prediction head, we simultaneously regress the pose of the CAD model that aligns it with the image observation.

A shared embedding space between image and CAD can be difficult to effectively construct due to the strong differences between the two domains. While mapping image observations of an object together with the full CAD models into a shared embedding space has shown promise [34, 30], this approach tends to struggle with generalization to views of objects without exact matches in the CAD database. Thus, rather than constructing an embedding space which maps similar image observations of an object together with full CAD models, we aim to learn an embedding space which captures not only global semantic similarity between image and CAD, but mid-level and low-level geometric similarities. We propose to learn the embedding of object parts and CAD parts by constructing a shared feature space where patches of object images lie close to similar patches of CAD objects. This enables reasoning about similar parts for retrieval in a scenario without exact CAD matches for a new image view, enabling more robust CAD reconstruction.

3.2. 2D Object Detection

We leverage a state-of-the-art 2D object detection and instance segmentation backbone to inform our 3D shape reasoning. From an input RGB image, 2D object bounding boxes and class labels are localized using RetinaNet [36], and instance segmentation masks are predicted using ShapeMask [31]. The learned features from the 2D object detection guide our shape prediction; for a detected object k, we use the predicted box for an object to crop features f_k corresponding to the object, and multiply with the instance mask prediction m_k. We then use $m_k \circ f_k$ as input for the image-shape embedding as well as pose alignment.

3.3. Patch-Based Joint Embedding Learning

Our approach centers around constructing a patch-based joint embedding space between the two domains of image observations of an object and 3D CAD model representations of objects. While humans can establish perceptual correspondence between images and CAD models, it is challenging to bridge the domains due to strong differences in representation: in contrast to a 3D geometric CAD model, an image is view-dependent, colored, and contains lighting and material effects. Moreover, in real-world scenarios, we typically do not have exact CAD matches to the image views as groundtruth annotation. We construct an embedding space between image patches and CAD patches, to enable reasoning about mid-level and low-level structural

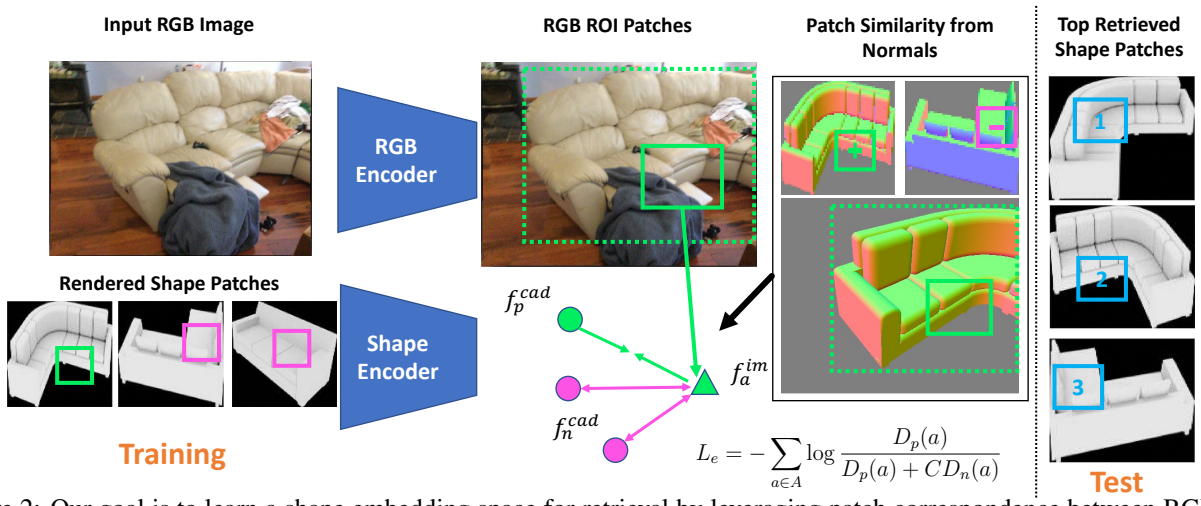

Figure 2: Our goal is to learn a shape embedding space for retrieval by leveraging patch correspondence between RGB and shape. At train time, we sample RGB patches from object regions and rendered shape patches from the object class. We shape the embedding space with a contrastive loss, and regularize the learning with surface normal matching so that positive patches have high geometric similarity, while negative patches come from non-matching shapes with low geometric patch similarity. This patch-wise construction establishes more robust correspondence for shape retrieval from images at test time.

similarities between objects, as many objects can share similarly structured parts while not matching exactly. Thus, we can establish part similarities where a global object mapping can struggle to fulfill a complete object match. By bridging the two domains in such a fashion, we can more easily recognize similar geometric structures in a new observation of an object not exactly represented in the CAD database.

Our patch-based joint image-shape embedding space is visualized in Figure 2, and is constructed based on patches of the image feature of an object and patches of rendered CAD models to n canonical views $\{c_0, ...c_{n-1}\}$ (similar to a light-field descriptor [8]). The representation of CAD models to their rendered views helps to reduce the domain gap between 2D image and 3D shape; we use $n = 16$ views with the canonical views determined by K-medoid clustering of the train views. To embed image and shape, we extract features from $m_k \circ f_k$ and sampled patches from the rendered CAD views using embedding modules composed of a series of 2D convolutions, resulting in f^{im} and $\{f_j^{\text{cad}}\}$, respectively. Each embedding network for image and CAD features is structured symmetrically, without shared weights as they operate on different domains. We then construct the patch embedding space by randomly sampling anchor patches from f^{im}, which we denote as f_a^{im}, and then establish positive and negative similarity with the $\{f_j^{\text{cad}}\}$. In all our experiments, we use a patch size of $1/3$ RGB-ROI or rendered shape image size. The embedding space is constructed with a noise contrastive estimation loss [43]:

$$L_e = -\sum_{a \in A} \log \frac{D_p(a)}{D_p(a) + CD_n(a)} \tag{1}$$

$$D_p(a) = \frac{1}{|P(a)|} \sum_{p \in P(a)} \exp\left(D(f_a^{\text{im}}, f_p^{\text{cad}})\right) \tag{2}$$

$$D_n(a) = \frac{1}{|N(a)|} \sum_{n \in N(a)} \exp\left(D(f_a^{\text{im}}, f_n^{\text{cad}})\right) \tag{3}$$

where L_e denotes the total loss, A the set of all anchor (query) patches, $P(a)$ and $N(a)$ the positive and negative matches for the query patch a, $C = 24$ a weighting value, $D(x, y) = (x/||x||)^T(y/||y||)/\tau$ with $\tau = 0.15$. $D_p(a)$ and $D_n(a)$ are the mean exponentiated weights of positive and negative pairs. To further improve learning efficiency, we exclude empty RGB and shape patches in the embedding loss, as determined by the rendered binary masks.

Our overall loss is similar to Mask2CAD, but because we operate on patch-level correspondence, we removed hard-positive mining due to relaxed constraints on patch matching vs. exact instance matches. The formulation of our loss is different from standard InfoNCE loss because we have multiple positives (shape rendering patch) for each query patch. Thus, we need to balance the ratio of positives/negatives per batch via the C parameter.

Patch similarity for embedding construction. To train our embedding construction, we establish patch similarity for matching and non-matching patches between image and CAD patches by estimating their geometric similarity. We use rendered normals from the CAD models (with normals represented in canonical space) and their corresponding patches to represent local geometry, and for the image, we use patches of the rendered normals of the ground truth corresponding CAD model. For each patch of normals, we

compute its descriptor by a self-similarity histogram, evaluated as the histogram over all pairwise angular distances of the normals in the patch; histograms are normalized to sum to 1. This allows us to estimate orientation-independent geometric similarity. We then measure the difference between two patches of normals by IoU of their self-similarity histograms. Positive matches to a query are determined by patches corresponding to a ground truth CAD annotation with self-similarity IoU $> \theta_p$, and negative matches by patches corresponding to non-corresponding CAD models with self-similarity IoU $< \theta_n$. Since a ground truth CAD annotation may contain patches that are dissimilar to the query patch and non-ground truth CAD models may contain patches that are similar to the query, we empirically found that double thresholds helped to avoid such associations. We set $\theta_p, \theta_n = 0.4, 0.6$ in our implementation.

We additionally employ hard negative mining by sampling the top negative patches by distance to the query. During training, we take $16\times$ the number of objects per image for hard negative examples. This enables for better distinguishing on difficult cases, and an improved embedding space. We set $|N(a)| = 1024$ for each anchor patch. Regarding hard-positive mining, we observe that it hurts the performance with a fixed top-K mining due to unstable number of positive pairs per batch. To remedy this, we average the weights of all positive pairs and treat them as one positive sample $D_p(a)$, which leads to significantly more stable learning and better performance.

3.4. Patch-Based Retrieval

Since our joint embedding of images and shapes is constructed patch-wise, we can leverage many patch retrievals for a more robust, comprehensive shape retrieval. We use randomly sampled patches from CAD renderings to construct the database for retrieval. Then for a detected object in an image, we randomly sample K_q patches from f^{im}, and for each patch, we retrieve K_r patches from the database. The K_r retrieved patches are then used to decide the corresponding CAD model of the patch query by majority voting, resulting in K_q CAD models for each patch; the final shape retrieval is obtained by majority vote of the K_q patch-retrieved CADs, excluding those retrieved by patches fully outside the predicted instance mask. While an image-CAD mapping based on full image view of the object and whole CAD model might struggle to retrieve from a global similarity perspective under inexact matches, our patch-based shape retrieval encourages the retrieved shapes to more comprehensively match the image.

3.5. Pose Prediction

We simultaneously predict the pose of the 3D shape corresponding to its 2D image observation in a separate branch. Similar to [30], we predict the rotation of the shape by a rotation classification followed by a regression refinement, and the translation as an offset from the 2D bounding box center. To obtain the estimated rotation, use rotation bins computed by K-medoid clustering of the train object rotations as quaternions, and predict the bin using a cross entropy loss, followed by predicting a refinement offset quaternion with a Huber loss [26]. Translation is estimated as an offset from the predicted bounding box center as a ratio of the box dimensions, and optimized with a Huber loss.

3.6. Implementation Details

Our ShapeMask [31] instance segmentation backbone (ResNet-50-FPN) is initialized with COCO pretraining, and our embedding for both image and CAD renderings uses a ResNet-18-FPN backbone with random intialization. We train our instance segmentation for amodal bounding box prediction instead of modal boxes in the standard COCO setup, as this can capture more consistent context and provides more stable guidance for the pose translation estimation. We additionally apply data augmentation to improve generalization, including HSV color jitter, ROI box jitter, and image scale jitter during training.

We train our approach for 36K iterations using a batch size of 256 on ScanNet, which takes $\approx$ 2 days. The learning rate is initialized to 0.16 and decreased by 10x at 24K and another 10x at 30K iterations. In terms of inference time, Patch2CAD takes $\approx$ 74ms per image [1], 58ms model + 16ms retrieval (vs. Mask2CAD $\approx$ 60ms), with unoptimized parallel patch retrievals.

4. Experiments

We evaluate our approach on the ScanNet dataset [11], which contains challenging real-world imagery of multiple objects per image in cluttered indoor environments, with many occlusions, partial views and varying lighting conditions. The ScanNet dataset contains 1513 indoor scenes; we use the Scan2CAD [2] annotations of ShapeNet [7] CAD models to the ScanNet scenes to provide ground truth CAD correspondences for training and evaluation. Note that there are no exact matches between the CAD models to the real-world imagery, as is reflective of many real-world application scenarios. Following the Mask2CAD [30] evaluation protocol on ScanNet, we use the 25K frame subset provided by the dataset for training and validation, containing 19387 train and 5436 validation images, respectively.

In addition, we evaluate our approach on the Pix3D dataset [52], which contains $10,069$ images of indoor furnitures labeled with corresponding CAD models. We use the train/test split by Mesh R-CNN [19] for direct comparison.

[1] Measured on Pix3D for comparison with Mask2CAD [30].

ScanNet 25K	AP	AP50	AP75	*bed*	*sofa*	*chair*	*cabinet*	*trashbin*	*display*	*table*	*bookshelf*
Mask2CAD [30]	8.4	23.1	4.9	14.2	**13.0**	13.2	**7.5**	7.8	5.9	2.9	3.1
Patch2CAD (Ours)	**10.3**	**26.0**	**6.6**	**18.8**	12.4	**17.6**	**7.5**	**8.6**	**10.8**	**3.3**	**3.3**

Table 1: Performance on ScanNet [11]. We report mean AP^{mesh} and per-category AP^{mesh}.

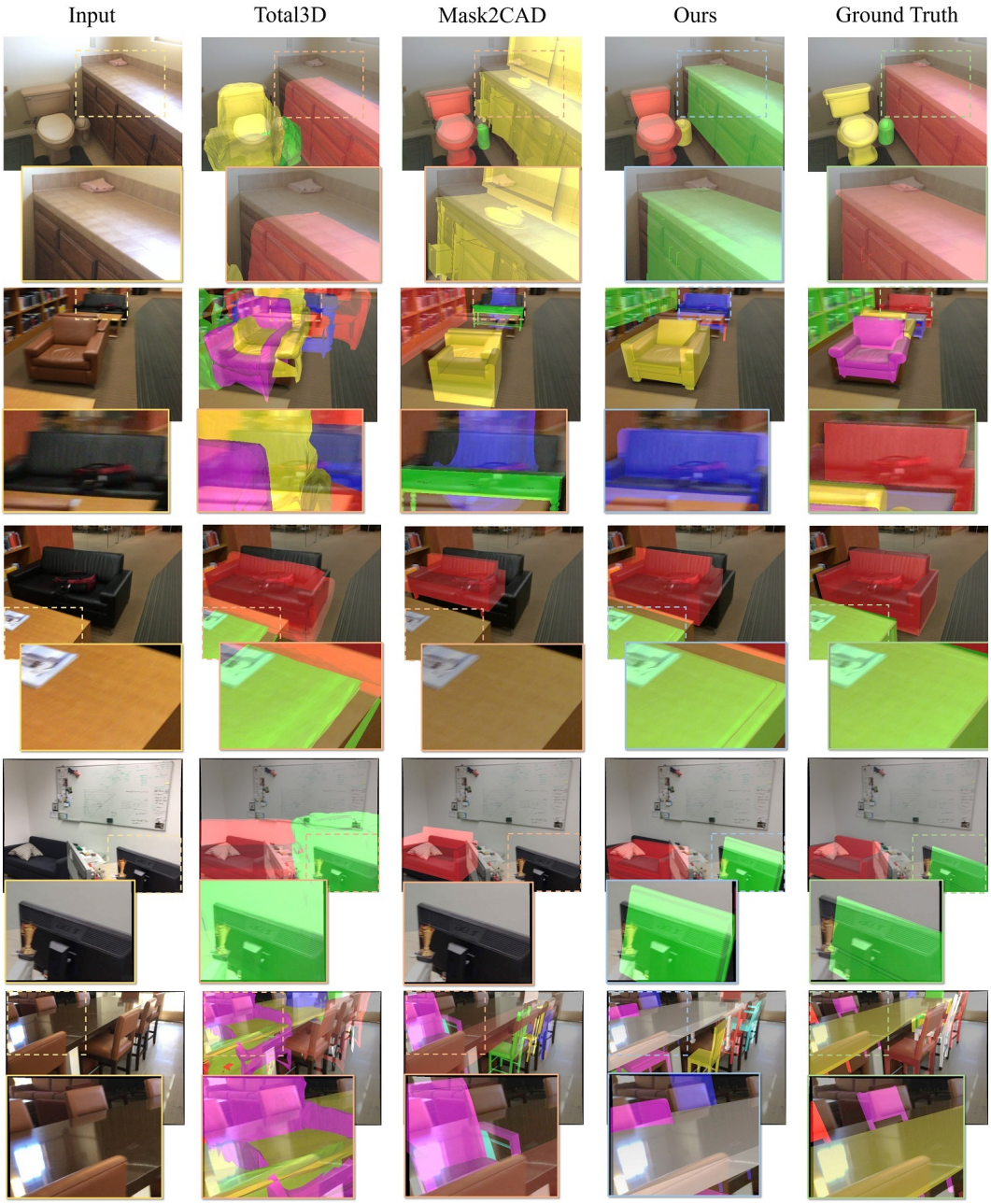

Figure 3: Qualitative results on ScanNet [11] images, in comparison with state of the art Total3D [42] and Mask2CAD [30]. Our patch-based shape embedding results in more accurate shape retrieval as well as more robust retrieval for strongly occluded objects (see rows 3, 4, 9, 10). Note that different colors denote distinct object instances in the visualization.

Evaluation metrics. We adopt previously established metrics for both 2D and 3D evaluation. For evaluating 2D outputs, we employ the predominant metrics from 2D object recognition: APbox and APmask on the 2D detections

ScanNet 25K	AP	AP50	AP75	*bed*	*sofa*	*chair*	*cabinet*	*trashbin*	*display*	*table*	*bookshelf*
Total3D [42]	1.4	6.3	0.2	1.9	4.3	1.5	0.8	0.1	0.0	0.7	2.1
Mask2CAD [30]	10.5	33.3	4.5	13.9	**13.1**	14.8	11.6	10.8	8.8	4.1	7.4
Patch2CAD (Ours)	**12.9**	**37.5**	**6.6**	**14.5**	11.6	**18.8**	**12.4**	**13.0**	**19.0**	**5.7**	**8.1**

Table 2: Performance on ScanNet [11] using groundtruth 2D detections. We report mean AP^{mesh} and per-category AP^{mesh}.

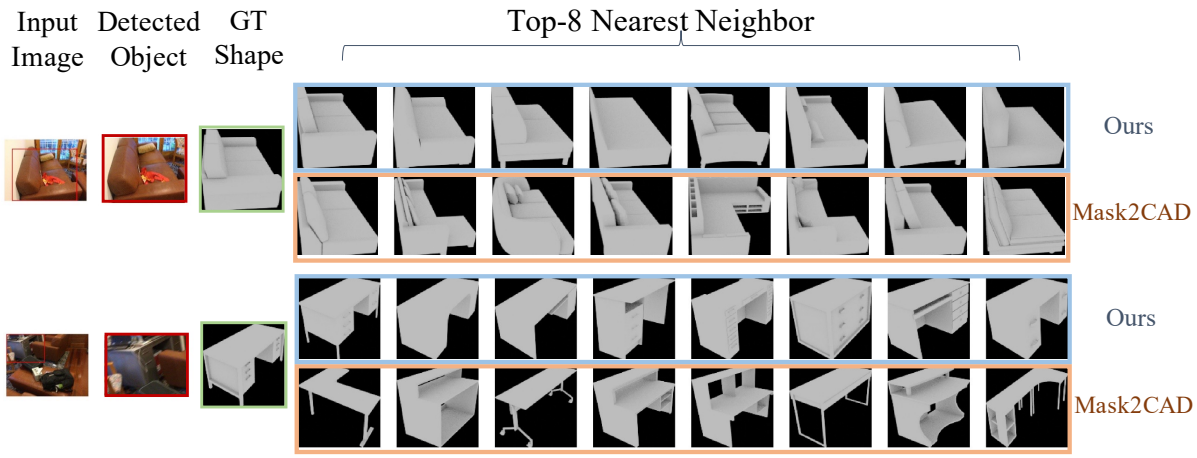

Figure 4: Top-k nearest neighbor retrievals from detected objects in ScanNet [11] images with Scan2CAD [2] ground truth CADs, in comparison to Mask2CAD [30] (same detection inputs). Our approach achieves a more consistent shape embedding space, enabling robust top-k retrieval with structurally similar CAD associations for not only the top-1 nearest neighbor.

Method	Patch Size	Normals	AP	AP50
Mask2CAD [30]	1.0		8.4	23.1
Patch2CAD (Ours)	1.0	V	9.4	24.7
Patch2CAD (Ours)	0.5	V	9.5	24.5
Patch2CAD (Ours)	**0.33**	V	**10.3**	**26.0**
Patch2CAD (Ours)	0.25	V	10.0	25.8

Table 3: Performance vs patch size and use of normals. 1.0 patch size corresponds to the full object size.

of objects. We use the newly introduced APmesh metric [19] to evaluate the 3D shape and pose predictions for the 3D objects. Similar to Mask2CAD [30], we evaluate APmesh metrics at IoU 0.5 (AP50) at IoU 0.75 (AP75), as well as AP as the mean over AP50-AP95, using 10 IoU thresholds between 0.5 and 0.95. For more consistent reproducibility, we report our evaluations as an average of 2 independent runs. The thresholds used in F-scores follow [30] on ScanNet and [19, 30] on Pix3D.

Comparison to the state of the art. In Table 1, we evaluate our 3D object understanding from a single image in comparison to Mask2CAD [30] on the ScanNet [11] benchmark proposed by Mask2CAD. Our improvement on Mesh AP50 is larger than AP75, showing that Patch2CAD maintains more robust shape estimate even when it retrieves non-exact matches. Mask2CAD also takes a retrieval-based approach, but maps full image observations of an object to entire CAD models, which tends towards overfitting and struggles with new test images whose objects do not exactly match the database; our patch-level embedding enables more robust retrieval and alignment by establishing correspondence with similar object parts rather than the more strict requirement of the full object. Additionally, this can help to retrieve and align objects that are occluded or partially visible in the input image (see Figure 3).

What is the effect of patch-wise embedding learning on representing shape geometry? In Table 1, we see that our patch-based embedding improves a retrieval-based 3D object reconstruction, in contrast to the whole-shape embedding of Mask2CAD. We additionally evaluate our patch-based embedding for retrieval given ground truth 2D detections in Table 2, showing consistent improvement over both the Mask2CAD retrieval and the mesh generation approach of Nie *et al.* [42]. Note that we use the training scheme of Nie *et al.* on SUN RGB-D [51], as they use scene layout information during training (similar to ScanNet, SUN RGB-D is also captured from real indoor scenes with a PrimeSense-based sensor). Finally, we evaluate Patch2CAD given ground truth 2D detections as well as pose (*i.e.*, evaluating shape only) in comparison with Nie *et al.* [42] as well as Mask2CAD [30] in Table 4, using an F-score for shape reconstruction evaluation. Even with a

ScanNet 25K	Mean	*bed*	*sofa*	*chair*	*cabinet*	*trashbin*	*display*	*table*	*bookshelf*
Total3D [42]	52.4	58.8	**72.6**	**69**	41.5	38.9	35.9	44.4	58.4
Mask2CAD [30]	60.6	63.1	64.4	66.1	**61.0**	68.3	58.7	47.1	56.3
Patch2CAD (Ours)	**63.8**	**64.3**	62.0	68.1	59.9	**71.6**	**73.9**	**51.9**	**58.9**

Table 4: Mean F-score and category F-score on ScanNet [11] using groundtruth 2D detections and evaluating shape only.

Pix3D $\mathcal{S}_1$	AP	AP50	AP75
Mesh R-CNN [19]	17.2	51.2	7.4
Mask2CAD [30]	**33.2**	**54.9**	**30.8**
Patch2CAD	30.9	51.7	28.2

Table 5: Performance on Pix3D [52] $\mathcal{S}_1$. We report mean AP^{mesh} following [19, 30].

shape-only prediction, strong occlusions in the image views can be challenging; Patch2CAD maintains more robustness.

Effect of patch size and use of surface normals. Table 3 analyzes various patch sizes and with/without surface normals. The first row corresponds to Mask2CAD; ours on the fourth row. Our $\frac{1}{3}$ patch basis and use of normals helps to notably improve shape retrieval.

Comparison on Pix3D. Table 5 shows a comparison with both generative and retrieval methods on Pix3D [52]. Pix3D presents a scenario with exact shape matches in simpler scenes than ScanNet. Patch2CAD performs notably better than Mesh R-CNN [19], and is competitive with Mask2CAD [30], whose full object matching approach well-suits the scenario with exact 3D matches.

How does a patch-wise embedding mold the space for top-k retrieval? We evaluate our patch-based image-CAD embedding space by analyzing the top-k nearest neighbor CAD models retrieved for a given detected object on the ScanNet validation set. We visualize the top-8 retrieved CAD models for various image object detections in Figure 4. In contrast to the full-shape mapping of image-CAD established by Mask2CAD, our patch-wise embedding construction encourages more similarly structured CAD shapes to be voted for by the patches, resulting in geometrically consistent top-k.

Quantitatively, we analyze our top-k shape retrieval by evaluating the recall from the k retrieved shapes. We compare with the state-of-the-art Mask2CAD [30] approach in Figure 5, using $k = 1$ to 24. Our patch-based approach consistently produces more accurate shape retrieval.

Limitations. While our Patch2CAD approach demonstrates a more robust joint embedding space construction between images and CAD models, there are various directions for development. For instance, our patch-based retrieval can produce more robust CAD retrieval results, but cannot represent shapes that differ significantly from the database; we believe a part-based synthesis or deformation approach from our various patch retrievals holds promise. Additionally, our approach tackles object shape and structure, but does not represent the full scene geometry, which is an important direction towards comprehensive 3D perception.

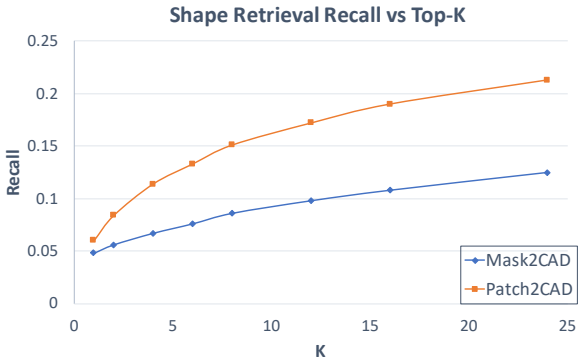

Figure 5: Shape retrieval comparison with Mask2CAD [30]

5. Conclusion

In this paper, we present Patch2CAD, which establishes patch-based correspondence between 2D images and 3D CAD models for a robust construction of a shared embedding space to map between the two domains. This enables CAD-based understanding of the shapes of objects seen from a 2D image, representing each object as a posed, lightweight, complete mesh. We demonstrate that our patch-wise embedding learning can construct a more meaningful embedding space for nearest neighbor retrieval, and more robust shape estimation for complex real-world imagery under many occlusions. We believe that this brings understanding forward in bridging these domains of 2D-3D as well as real-synthetic, which opens avenues in domain transfer, content creation, and 3D scene understanding.

Acknowledgements

We would like to thank our colleagues at Google Research for their advice, and the support of the Bavarian State Ministry of Science and the Arts as coordinated by the Bavarian Research Institute for Digital Transformation (bidt) for Angela Dai.

References

[1] Mathieu Aubry, Daniel Maturana, Alexei A Efros, Bryan C Russell, and Josef Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3762–3769, 2014. 3

[2] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X. Chang, and Matthias Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. *CVPR*, 2019. 2, 3, 5, 7

[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 International Conference on Computer Vision*, pages 2551–2560, 2019. 3

[4] Armen Avetisyan, Tatiana Khanova, Christopher Choy, Denver Dash, Angela Dai, and Matthias Nießner. Scenecad: Predicting object alignments and layouts in rgb-d scans. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 596–612. Springer, 2020. 3

[5] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5965–5974, 2016. 3

[6] Thomas O Binford. Survey of model-based image analysis systems. *The International Journal of Robotics Research*, 1(1):18–64, 1982. 3

[7] 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. 3, 5

[8] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3d model retrieval. In *Computer graphics forum*, volume 22, pages 223–232. Wiley Online Library, 2003. 4

[9] Roland T Chin and Charles R Dyer. Model-based recognition in robot vision. *ACM Computing Surveys (CSUR)*, 18(1):67–108, 1986. 3

[10] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016. 2

[11] 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. 2, 5, 6, 7, 8

[12] Angela Dai and Matthias Nießner. Scan2mesh: From unstructured range scans to 3d meshes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5574–5583, 2019. 2

[13] Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, and Andrea Tagliasacchi. Cvxnet: Learnable convex decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31–44, 2020. 2

[14] Maximilian Denninger and Rudolph Triebel. 3d scene reconstruction from a single viewport. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[15] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019. 2

[16] 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. 2

[17] Georgios Georgakis, Srikrishna Karanam, Ziyan Wu, and Jana Kosecka. Learning local rgb-to-cad correspondences for object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8967–8976, 2019. 3

[18] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1, 2

[19] 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. 2, 5, 7, 8

[20] Alexander Grabner, Peter M Roth, and Vincent Lepetit. Location field descriptors: Single image 3d model retrieval in the wild. In *2019 International Conference on 3D Vision (3DV)*, pages 583–593. IEEE, 2019. 3

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 1, 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[24] Qixing Huang, Hai Wang, and Vladlen Koltun. Single-view reconstruction via joint analysis of image and shape collections. *ACM Transactions on Graphics (TOG)*, 34(4):1–10, 2015. 3

[25] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3D scene parsing and reconstruction from a single RGB image. In *European Conference on Computer Vision*, pages 194–211. Springer, 2018. 3

[26] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992. 5

[27] Hamid Izadinia and Steven M Seitz. Scene recomposition by learning-based icp. In *CVPR*, 2020. 3

[28] Hamid Izadinia, Qi Shan, and Steven M Seitz. Im2cad. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5134–5143, 2017. 3

[29] Young Min Kim, Niloy J Mitra, Qixing Huang, and Leonidas Guibas. Guided real-time scanning of indoor objects. In *Computer Graphics Forum*, volume 32, pages 177–186. Wiley Online Library, 2013. 3

[30] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Mask2CAD: 3D shape prediction by learning to segment and retrieve. In *Eur. Conf. Comput. Vis.*, 2020. 2, 3, 5, 6, 7, 8

[31] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9207–9216, 2019. 2, 3, 5

[32] Hei Law and Jia Deng. Objects as paired keypoints. *ECCV*, 2018. 2

[33] Yangyan Li, Angela Dai, Leonidas Guibas, and Matthias Nießner. Database-assisted object retrieval for real-time 3d reconstruction. In *Computer Graphics Forum*, volume 34, pages 435–446. Wiley Online Library, 2015. 2, 3

[34] 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. 3

[35] J.J. Lim, H. Pirsiavash, and A. Torralba. Parsing ikea objects: Fine pose estimation. *ICCV*, 2013. 3

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2, 3

[37] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[39] Francisco Massa, Bryan C Russell, and Mathieu Aubry. Deep exemplar 2d-3d detection by adapting from real to rendered views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6024–6033, 2016. 3

[40] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2

[41] 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. 2

[42] 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. 2, 6, 7, 8

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[44] 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 Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[45] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 2

[47] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3577–3586, 2017. 2

[48] Lawrence G Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963. 3

[49] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2

[50] Tianjia Shao, Weiwei Xu, Kun Zhou, Jingdong Wang, Dongping Li, and Baining Guo. An interactive approach to semantic modeling of indoor scenes with an rgbd camera. *ACM Transactions on Graphics (TOG)*, 31(6):1–11, 2012. 3

[51] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 7

[52] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3D: Dataset and methods for single-image 3D shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2974–2983, 2018. 2, 3, 5, 8

[53] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2088–2096, 2017. 2

[54] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2019. 3

[55] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018. 2

[56] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1042–1051, 2019. 2

[57] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in neural information processing systems*, pages 82–90, 2016. 2

[58] Yue Wu, Yinpeng Chen, Zicheng Liu Lu Yuan, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization for object detection. *CVPR*, 2020. 2

[59] 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 International Conference on Computer Vision*, pages 4541–4550, 2019. 2