Learning Geometric-aware Properties in 2D Representation Using Lightweight CAD Models, or Zero Real 3D Pairs

Pattaramanee Arsomngern   Sarana Nutanong   Supasorn Suwajanakorn
VISTEC, Thailand
{pattaramanee.a_s19, snutanon, supasorn.s}@vistec.ac.th

Abstract

Cross-modal training using 2D-3D paired datasets, such as those containing multi-view images and 3D scene scans, presents an effective way to enhance 2D scene understanding by introducing geometric and view-invariance priors into 2D features. However, the need for large-scale scene datasets can impede scalability and further improvements. This paper explores an alternative learning method by leveraging a lightweight and publicly available type of 3D data in the form of CAD models. We construct a 3D space with geometric-aware alignment where the similarity in this space reflects the geometric similarity of CAD models based on the Chamfer distance. The acquired geometric-aware properties are then induced into 2D features, which boost performance on downstream tasks more effectively than existing RGB-CAD approaches. Our technique is not limited to paired RGB-CAD datasets. By training exclusively on pseudo pairs generated from CAD-based reconstruction methods, we enhance the performance of SOTA 2D pre-trained models that use ResNet-50 or ViT-B backbones on various 2D understanding tasks. We also achieve comparable results to SOTA methods trained on scene scans on four tasks in NYUv2, SUNRGB-D, indoor ADE20k, and indoor/outdoor COCO, despite using lightweight CAD models or pseudo data. Please visit our page: https://GeoAware2dRepUsingCAD.github.io/

1. Introduction

Recent 2D visual representation learning approaches, such as contrastive learning [3, 6, 10, 17, 28] or masked autoencoder [20], are widely used to tackle various problems in computer vision due to their ability to encode rich visual features. While these methods have shown exceptional results on 2D image classification, they still have shortcomings in other 2D understanding tasks that involve instance-level reasoning. Prior research [25] also shows that models pre-trained using image augmentations [8, 21] or supervised labels [13] could not deliver satisfactory results when applied to downstream tasks such as semantic segmentation, instance segmentation, and object detection [12, 42].

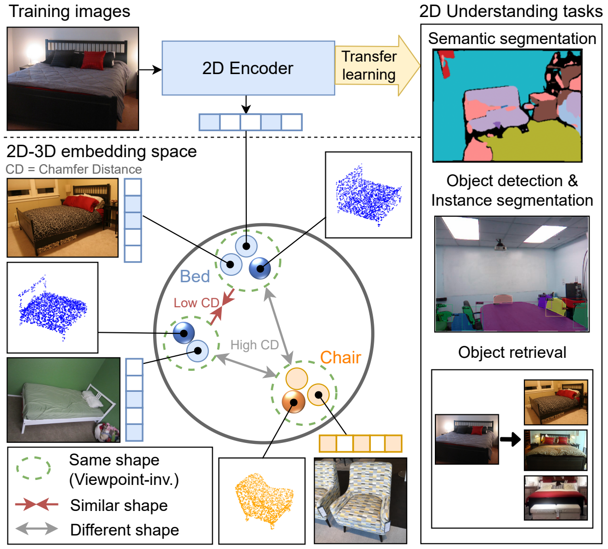

Figure 1. **Overview concept of our solution.** We leverage CAD models to train a joint 2D-3D space such that images of objects with similar shapes, based on the *Chamfer distance*, are attracted to each other, while images with different shapes are separated. This results in a continuous geometric-aware space where the distance between two points reflects their geometric similarity, which could be utilized for downstream 2D object understanding tasks.

To alleviate this, Hou et al. [25] proposed to learn 3D geometric priors, such as view-invariance, from 3D data and transfer the learned priors to 2D representations. In particular, their model is first pre-trained on ScanNet [12], a database of multi-view RGB-D scans, using contrastive learning and later used as initialization for fine-tuning networks on downstream tasks. Chen et al. [5] further extend this work by utilizing additional priors through learning to group nearby points that refer to the same object part from 3D scenes.

The key concept of their introduced new paradigm is to share useful 3D priors from 3D data with 2D representations. However, previous studies have investigated mostly 3D priors related to viewpoint invariance from 3D scene datasets, which are often limited in size and scene variations due to the laborious data collecting and labeling process [24]. This scarcity of large-scale 3D data also limits the number of available 3D priors for learning the invariance, hence hindering further performance gains. This paper questions whether there are other effective 3D priors for 2D downstream tasks and whether they can be learned from other forms of 3D data that are lighter and easier to obtain.

We begin our exploration by considering how to learn an embedding space that maps together images with the same or similar geometries. One solution is to learn from images that share the same 3D model, which inspires our interest in CAD models. Unlike 3D scenes, CAD models are lightweight, publicly available, and can be easily aligned with RGB images via web scraping or human annotation [2]. There exist studies that jointly learn 2D and 3D CAD representation for other CAD-related tasks, e.g., 3D classification [1, 27]. However, their 2D features, which are derived from augmentation-based CAD features, are insufficient to be directly applied to 2D object understanding tasks—they can group images with similar geometries but struggle to learn the distinctions between object categories.

We argue that useful *geometric-aware* representations should account for both similarities and differences in object geometry. Our key idea is to acquire such features by imitating the Chamfer distance between 3D objects in our embedding space and inducing this derived geometric awareness in our learned 2D representation. In particular, we learn our space by attracting the encoded features of geometrically similar CAD models in the mini-batch based on the Chamfer distance and repelling those with lower similarities. In contrast to other methods trained on supervised discrete signals like object labels or through 3D augmentations, our method produces a continuous 3D space that better captures the similarity and difference in geometry (see Section 5.1). In addition, we employ augmentation-based contrastive learning [6] to learn other useful visual feature properties, such as translation and color invariances. This results in a 2D representation in a 2D-3D space that contains rich visual information and strong geometric-aware properties, as shown in Fig. 1, which can be leveraged to improve 2D object understanding tasks.

To match our geometric-aware CAD features with corresponding 2D features, a paired RGB-CAD dataset, such as Pix3D [45], is required. However, by leveraging recent techniques [19, 31] that can reconstruct a CAD model from an input image, it is possible to generate *pseudo* CAD models for any images and use them to learn our method without a paired dataset. Adapting this pseudo-pair generation to other techniques that rely on scene scans is significantly harder, as synthesizing full 3D scenes with reasonable detail remains harder than reconstructing individual objects.

We demonstrate the effectiveness of our geometric-aware 2D representation in Fig. 2. Our features can group and differentiate objects based on their categories or subcategories, leading to improved performance on multiple 2D object understanding tasks using both ResNet-50 [23] and ViT-B [29] backbones. Our method trained on a pseudo-pair dataset also yields superior results over DINO [3] and MAE [20]. Remarkably, we also surpass a state-of-the-art method, Pri3D [25] without using any 3D scene scans in the following tasks: (i) semantic segmentation using NYUv2 [42] and indoor ADE20k [54]; (ii) object detection and instance segmentation using NYUv2 and in/outdoor COCO [35]; (iii) object retrieval using Pix3D [45].

To summarize, our contributions are as follows.

- We present a simple yet effective approach to inducing geometric-aware properties in 2D representation using lightweight CAD models. These can be either ground truth from RGB-CAD datasets or generated pseudo CAD pairs based on 2D-only data.
- We propose training objectives to learn a 2D-3D embedding space where feature similarity reflects geometric similarity based on the Chamfer distance.
- We enhance the performance of SOTA 2D representation learning techniques on four 2D object understanding tasks and achieve competitive results to SOTA that require 3D scene scans across five datasets, in both settings that use real or pseudo-RGB-CAD datasets.

2. Related works

2D Representation Learning. Contrastive learning [49] has continuously shown improvements in various downstream tasks. Its main concept is to maximize a similarity score between two different views (e.g., two different augmentations of the same instance [6, 9, 10, 17, 21], two instances with the same label [28], or two different encoders [3, 7]). Meanwhile, learning representation in autoencoder fashion [20] also has recently demonstrated a performance boost from contrastive learning works.

This paper shows that these 2D representation learning methods could not provide geometric-aware properties, which are critical in object understanding tasks.

3D Representation Learning requires specific network architecture for extracting features based on an input's data structure (e.g., rendered images [44], point clouds [39, 46], or meshes [15, 16]). These architectures can encode geometric information of each 3D shape into representations with a deep understanding of shape structure. Later, contrastive learning in 3D, such as point cloud augmentation [52], or multiple modalities [1, 53], showed their improvement in 3D classification, retrieval, and part segmentation problems. Concurrently, 3D reconstruction techniques [11, 14, 37] also deliver promising results in similar tasks.

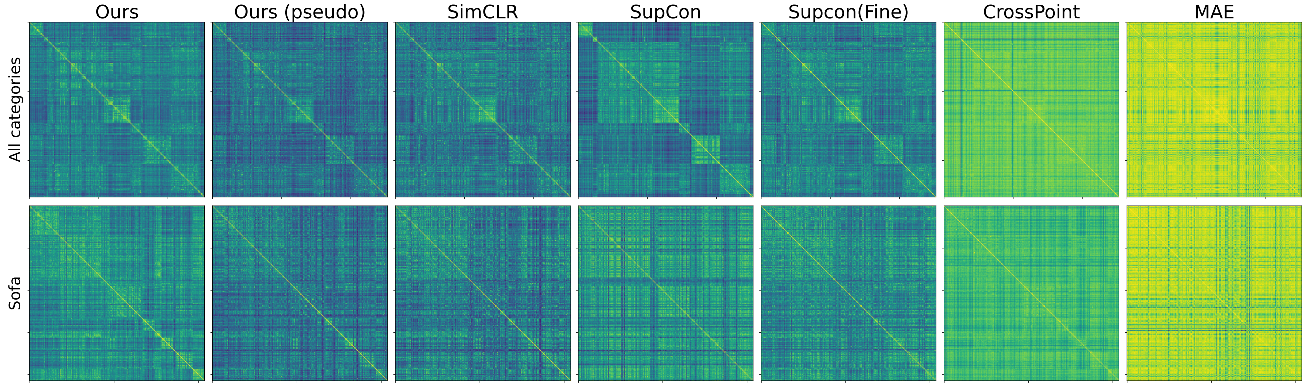

Figure 2. Visualization of the pairwise cosine similarity between the learned representations of objects from each method. The bright color indicates higher similarity. The first row shows the similarity scores of all validation images in Pix3D [45] dataset, sorted by object category and subcategory (i.e., object models). The second row zooms in on a single category (Sofa) from the first row. Our method shows a better grouping of the same sofa than others, especially SimCLR [6] and SupCon [28], which can hardly differentiate between each sofa type. CrossPoint [1] and SupCon (fine), unlike others, can group the same sofa but fail to separate sofas from other categories in the first row. Surprisingly, MAE [20] also delivers visual features without subcategory awareness. Compared to SimCLR, Ours (pseudo) provides distinguishable features (with darker colors) between each sofa type. Additional studies on Ours (pseudo) are provided in Appendix B.

In this paper, we adapt the encoded 3D features with rich geometric information to share the ability to recognize geometric similarities in each object to 2D representations.

Multimodal Representation Learning gains attraction due to its ability to share modality-specific contexts. CLIP [40] facilitates various vision-language tasks by contrastive learning on text-image data. Similarly, 2D-3D representation learning transfers geometric information from 3D to 2D features [5, 25, 36]. Pri3D [25] introduces a contrastive learning method on multi-view RGB frames and 3D scene scans [12]. These modal pairs allow them to learn view and 3D priors in 2D features and achieve impressive 2D indoor scene understanding performance. Later, Set-InfoNCE [5] improves upon Pri3D by fine-tuning the Pri3D pre-trained model using additional 3D priors, e.g., sets of nearby points referring to the same object part, generated from training scans. While these studies have focused on scene scans, this paper demonstrates that geometric priors from CAD models can achieve comparable or better object understanding than learning on 3D scenes.

Some studies, e.g., CrossPoint [1, 27], have been conducted to share visual context to guide CAD features (2D → 3D). Their learned 2D features, on the other hand, fail to discriminate object categories, resulting in poor object understanding results. We then investigate a more efficient method of learning and sharing modality-specific context from CAD to RGB without hurting 2D task performance.

2D-3D Object Matching has been studied for searching or generating a 3D model of objects appearing in a given image. [19, 30–32] use existing CAD databases to train an RGB-CAD retrieval and alignment module in a multi-task learning fashion using a given set of ground truths (e.g., camera intrinsic, depth map, mask, or pose parameters). [26, 34] utilize category labels to guide a joint RGB-CAD space for 2D-3D retrieval tasks. Another approach is to reconstruct a 3D model from the image directly [16, 18, 47]. However, the generated 3D may be less realistic than retrieving human-designed 3D from the databases.

We utilize these works to generate CAD pairs for our pre-training RGB images. Learning representation on just pseudo-RGB-CAD pairs can surpass SOTA competitors.

3. Proposed method

Our goal is to construct a pre-trained model that can provide a 2D representation $\mathbf{z}_i^{\mathrm{I}}$ of an input image I_i with useful geometric priors to enhance 2D object understanding. Unlike prior works [5, 25], we aim to learn the representations *without* 3D scene scans using other efficient alternatives.

We propose a novel approach that can obtain such geometric priors from a lightweight and publicly available type of 3D data in the form of CAD models, while retaining discriminative representations for 2D tasks. Given an image I_i and its associated CAD model G_i, our solution constructs a 2D-3D embedding space with strong geometric-aware properties by learning CAD features $\mathbf{z}_i^{\mathrm{G}} = f^{\mathrm{G}}(G_i)$ along with rich visual information from 2D features $\mathbf{z}_i^{\mathrm{I}} = f^{\mathrm{I}}(I_i)$. We show that the acquired 2D representations, where their feature similarity reflects geometric similarity, are more effective than solely relying on standard augmentation invariant CAD features and perform almost as effectively as 3D priors from scene scans on common 2D understanding tasks. The overview of our proposed framework is shown in Fig. 3.

An extension to training our model on generated RGB-CAD pairs is further introduced in order to learn geometric-aware representations using *just* RGB training data.

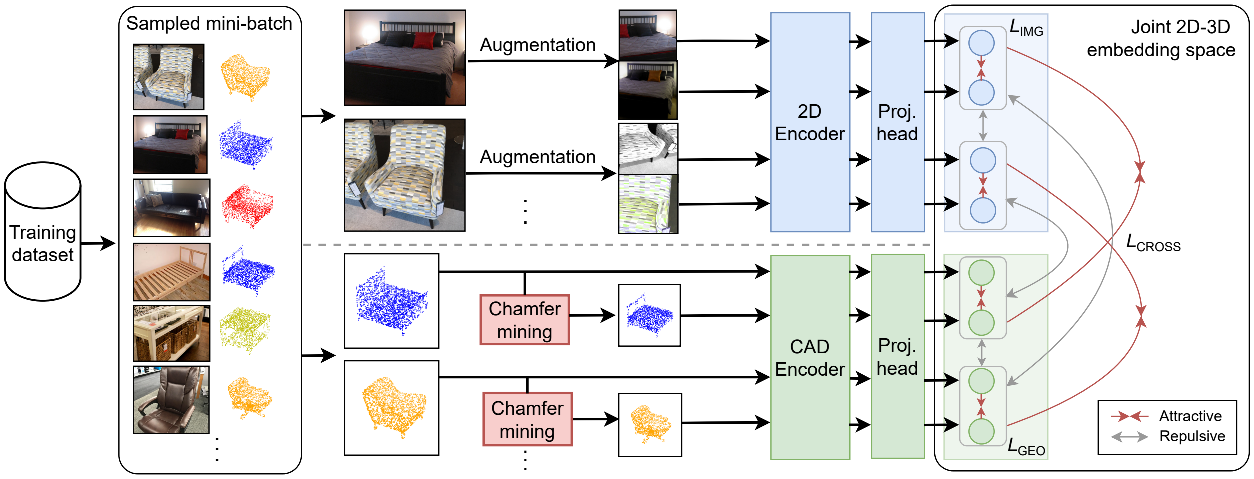

Figure 3. **Our pre-training strategy.** We learn 2D representations on a joint 2D-3D space from RGB-CAD pairs based on three loss functions. L_{GEO} focuses on learning CAD features from two similar CAD models mined from Chamfer distance, L_{IMG} focuses on learning visual differences between two image augmentations, and L_{CROSS} shares geometric awareness from CAD features to 2D representation.

3.1. Training objectives

The core idea of our training objective is to share two modality-specific contexts, i.e., strong geometric-aware properties from $\mathbf{z}_i^{\text{G}} \in \mathbf{z}^{\text{G}}$ and rich visual information from $\mathbf{z}_i^{\text{I}} \in \mathbf{z}^{\text{I}}$, in the joint 2D-3D embedding space by training on RGB-CAD pairs (I, G). To achieve this objective, we use three loss functions: L_{GEO}, L_{IMG}, and L_{CROSS}. The first loss L_{GEO} helps the 3D encoder $f^{\text{G}}(\cdot)$ construct a geometric-aware embedding space, such that CAD models in G with similar shapes, based on the Chamfer distance, are located near each other. The second loss L_{IMG} helps the 2D encoder $f^{\text{I}}(\cdot)$ learn useful visual representation properties, e.g., translation and color invariance from I. The third loss L_{CROSS} enforces the consistency between $\mathbf{z}_i^{\text{G}}$ and $\mathbf{z}_i^{\text{I}}$ for each (I_i, G_i) pair. Finally, the multi-task loss function is

$$L = L_{\text{GEO}} + L_{\text{IMG}} + L_{\text{CROSS}}. \tag{1}$$

Detailed explanations of the three components are given in the following subsections.

3.1.1 Geometric-aware CAD features

We first explain how L_{GEO} provides geometric awareness to our features. Given (I_i, G_i), we use G_i to train $f^{\text{G}}(\cdot)$ for extracting a geometric context $\mathbf{z}_i^{\text{G}}$ of each pair. We process each G_i by converting it to a point cloud and use DGCNN [46] as $f^{\text{G}}(\cdot)$ to capture local and global point cloud structure information, resulting in an encoded feature $\mathbf{z}_i^{\text{G}}$.

To learn the space, we follow a common contrastive learning strategy by maximizing the similarity between $\mathbf{z}_i^{\text{G}}$ and the encoded feature $\mathbf{z}^+$ of each positive term in a mini-batch. However, instead of selecting $\mathbf{z}^+$ based on object labels or augmentations as is commonly done, we select $\mathbf{z}^+$ based on a geometric-based similarity function, the Chamfer distance.

We calculate pairwise Chamfer distances across a mini-batch and select K point clouds with the lowest distances to be the positive terms for each G_i in the loss function. We call this process Chamfer mining. The set of encoded positive terms is referred to as $P(\mathbf{z}_i^{\text{G}})$. Finally, the contrastive loss is given as

$$L_{\text{GEO}} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{\mathbf{z}^+\in P(\mathbf{z}_i^{\text{G}})} -\log \frac{\exp(\mathbf{z}_i^{\text{G}}\cdot\mathbf{z}^+/\tau)}{\sum_{\mathbf{z}^-\in\mathbf{z}^{\text{G}}\setminus\{\mathbf{z}_i^{\text{G}}\}}\exp(\mathbf{z}_i^{\text{G}}\cdot\mathbf{z}^-/\tau)}, \tag{2}$$

where N is the batch size and τ is a tunable temperature parameter. Using $\mathbf{z}^+$ based on the Chamfer distance better captures the similarity and difference in 3D geometry and benefits 2D representation than other choices, e.g., 3D augmentation or category labels, which will be discussed later in Section 5.5 and Appendix B.

3.1.2 Learning to discriminate 2D visual features

Translation and color invariance are crucial in many vision problems [17]. Learning them along with geometric-aware properties potentially yields better results.

To acquire these invariances, we define the second function L_{IMG} by following a common practice in 2D representation learning. In particular, we use SimCLR [6] and learn to discriminate between two different augmentations of I_i (i.e., I_i and I_i^+):

$$L_{\text{IMG}} = l_{\text{IMG}}(\mathbf{z}^{\text{I}}, \mathbf{z}^{\text{I+}}) + l_{\text{IMG}}(\mathbf{z}^{\text{I+}}, \mathbf{z}^{\text{I}}),$$

$$l_{\text{IMG}}(\mathbf{z}^{\text{I}}, \mathbf{z}^{\text{I+}}) = \frac{1}{2N}\sum_{i=1}^{N} -\log \frac{\exp(\mathbf{z}_i^{\text{I}}\cdot\mathbf{z}_i^{\text{I+}}/\tau)}{\sum_{\mathbf{z}^-\in\mathbf{z}^{\text{I}}\cup\mathbf{z}^{\text{I+}}\setminus\{\mathbf{z}_i^{\text{I}}\}}\exp(\mathbf{z}_i^{\text{I}}\cdot\mathbf{z}^-/\tau)}, \tag{3}$$

where $\mathbf{z}_i^{\text{I+}} \in \mathbf{z}^{\text{I+}}$ is an encoded feature of I_i^+.

3.1.3 Cross-modal sharing the properties on the joint 2D-3D space

The third function L_{CROSS} aligns the geometric-aware features $\mathbf{z}^{\text{G}}$ with the 2D features $\mathbf{z}^{\text{I}}$ by constructing a common 2D-3D space for the two encoders, $f^{\text{G}}(\cdot)$ and $f^{\text{I}}(\cdot)$. To this end, we use cross-modal contrastive learning loss as a training objective as follows:

$$\begin{aligned} L_{\text{CROSS}} &= l_{\text{CROSS}}(\mathbf{z}^{\text{I}}, \mathbf{z}^{\text{G}}) + l_{\text{CROSS}}(\mathbf{z}^{\text{G}}, \mathbf{z}^{\text{I}}), \\ l_{\text{CROSS}}(\mathbf{z}^{\text{I}}, \mathbf{z}^{\text{G}}) &= \frac{1}{2N} \sum_{i=1}^{N} -\log \frac{\exp(\mathbf{z}_i^{\text{I}} \cdot \mathbf{z}_i^{\text{G}} / \tau)}{\sum_{\mathbf{z}^{-} \in \mathbf{z}^{\text{G}} \setminus \{\mathbf{z}_i^{\text{G}}\}} \exp(\mathbf{z}_i^{\text{I}} \cdot \mathbf{z}^{-} / \tau)}. \end{aligned} \tag{4}$$

Sharing $\mathbf{z}^{\text{G}}$ from L_{GEO} with $\mathbf{z}^{\text{I}}$ from L_{IMG} on L_{CROSS} using L results in a $f^{\text{I}}(\cdot)$ that can produce 2D representation with useful geometric and visual priors that show superior fine-tuning performance in the experimental section.

3.2. Learning geometric-aware representation on non-paired RGB datasets

Training a model to induce geometric-aware properties in 2D representation using Eq. (2) and Eq. (4) requires RGB-CAD pairs (I, G). However, obtaining such paired datasets is a cumbersome problem in every cross-modal learning framework. To address this, we propose to utilize 2D-3D object matching techniques, e.g., [16, 19, 30, 31, 47], to generate *pseudo-pair* of images and its associated 3D models, (I_i, G'_i), which can be used for training instead of the real pairs. Our experiments show that these pseudo pairs still help induce geometric-aware properties in our representations, resulting in better performance on 2D object understanding tasks, compared to state-of-the-art 2D representation learning approaches.

4. Experimental setup

4.1. Pre-training details

We use ResNet-50 [23] or ViT-B [29] initialized with supervised pre-trained ImageNet weights for the 2D encoder and use DGCNN [46] for the 3D encoder. We choose $\tau = 0.07$ for all equations and $K = 3$ in Eq. (2). Modality-specific projection heads are used for both 2D and 3D encoders to produce feature vectors z^{IMG} and z^{GEO} used in Eq. (1). Each head consists of two linear layers of sizes 1024 and 512, where the first one has ReLU activation. We pre-trained the model on the NVIDIA A100 GPU with a batch size of 256, then neglected the projection heads and retained only the 2D encoder for fine-tuning with downstream tasks. More details are given in Appendix A.

For the pre-train dataset, we use Pix3D [45] with the S1 train-test split [16]. This dataset consists of 10,069 paired RGB images of indoor scene furniture (e.g., bed, chair, sofa) and corresponding CAD models (subcategories) with a total of 395 shapes. We preprocessed each CAD model by uniform sampling random points to generate point clouds consisting of $N = 1024$ and applied point translation, rotation, and jittering to them. For RGB images, we applied the same augmentation strategy as [6] by randomly cropping an image into a size of 224×224, then randomly applying horizontal flipping, color jittering, grayscaling, and Gaussian blurring to the images.

4.2. Pseudo-pair generation setup

We utilized a state-of-the-art 3D shape retrieval model, ROCA [19], to generate a pseudo CAD model for any RGB image in indoor scene domains. ROCA was trained on Scan2CAD dataset [2] containing RGB-D frames of indoor scenes and annotated CAD models for each object in each frame. The Mask-RCNN [22] is employed to detect objects in an input image and generate bounding boxes of those detected ones. The feature of each bounding box will be further used in the ROCA retrieval module for predicting its associated CAD model listed in the ShapeNet [4] database.

The pairing for each RGB input was selected based on a predicted CAD model of the largest detected bounding box. See Appendix G for more ablation studies on pair generation. Later in the experiment section, we compare our model trained on RGB images from Pix3D dataset [45] and pseudo CAD pairs generated by ROCA with competitors.

4.3. Baseline competitors

All competitors except SupImg and Pri3D were initialized using pre-trained ImageNet weights and subsequently trained on Pix3D to ensure evaluation fairness.

Supervised ImageNet Pre-training (SupImg) [13] represents the base performance of a pre-trained model without additional training on indoor scene datasets.

SimCLR [6] and SupCon [28] provide the results when learning 2D representation on augmented RGB images (SimCLR), object categories (SupCon), or subcategories (SupCon (Fine)) in a CNN backbone.

CrossPoint [1] shows the performance of existing CAD-RGB representation learning. Note that this work has not been studied for solving 2D object understanding tasks.

Pri3D [25] and Set-InfoNCE [5] are state-of-the-art geometric-aware 2D representation learning works trained on multi-view RGB frames and 3D scans of ScanNet [12] dataset. Note that Set-InfoNCE did not publicly share their code, so we can only compare the results to their reported statistics on ScanNet [12] and NYUv2 [42] datasets.

DINO [3] and MAE [20] are state-of-the-art 2D representation learning models based on Vision Transformer (ViT) [29] architectures.

5. Experimental results

5.1. Properties of our learned features

Evaluating geometric-aware properties. Fig. 2 visualizes the pairwise cosine similarities between each encoded feature of objects from the Pix3D validation split. The representations were encoded with the 2D encoder (without a projection head) of each approach.

Our representations have high similarity scores for images with the same geometry (intra-subcategory) and others with the same object category (intra-category), while none of the competitors could reflect two properties at the same time. More visualizations from other categories and other competitors are provided in Appendix B.

One interesting aspect is that our representations also well capture the differences across inter-category and inter-subcategory samples. We acquire better mean intra and inter-category similarity scores than the self-supervised baselines, as shown in Fig. 4. Furthermore, Ours (pseudo) yields similar findings, particularly for inter-category. An additional study in Appendix B also demonstrates that Ours (pseudo) has better discrimination in inter-subcategory features. These geometric-aware properties in our learned representation lead to improved 2D object understanding performance in further subsections.

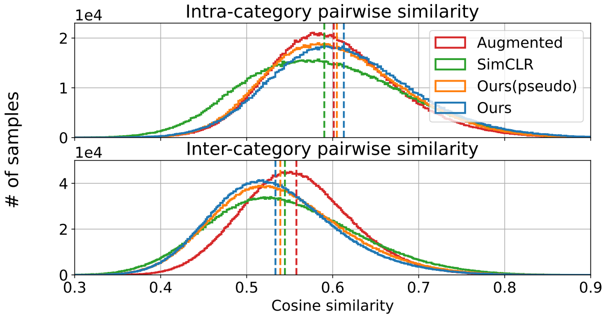

Figure 4. **Distribution of pairwise cosine similarity score among intra and inter-category samples.** Our representations learned by L_{GEO} based on the Chamfer distance have the highest mean score (vertical dashed lines) within intra-category samples and also have the lowest score in the inter-category scenario.

Object localization. We study whether our model can generalize to images with multiple objects, even though it was trained on a dataset of single objects. Fig. 5 shows that our learned features from the ViT-B backbone can be used for roughly localizing patches associated with object categories without fine-tuning. More details and results are provided in Appendix C. This property may arise from how our loss function encourages learning discriminative part features, which are useful for classifying, localizing, and segmenting multiple objects demonstrated in our subsequent experiments. This similar behavior is also observed in other 2D self-supervised models [3] trained on single-object datasets (e.g., ImageNet [13]).

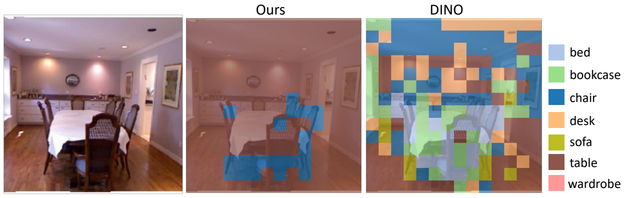

Figure 5. We compute the feature similarity between each patch in an unseen input image and the mean global feature of learned images in each category using models with a ViT-B backbone. *Without fine-tuning*, our model can identify the table and roughly localize patches associated with the chairs in the image by using the category of the most similar mean global feature of each patch.

5.2. 2D Semantic segmentation

Table 1 compares fine-tuning results in semantic segmentation tasks with other competitors. Following [5, 25], we train a U-Net [41] with residual connections for predicting segmentation masks with cross-entropy loss. The encoder part of U-Net is ResNet-50, while the decoder is conv layers with bi-linear interpolation. For ViT-B, we follow [20] by using UPerNet [51] as a segmentation head.

We employ two variants of mean intersection-over-union (mIoU) for the evaluation metric. The first mIoU, initially implemented by Pri3D [25], does not include pixels that were incorrectly predicted to be ignored classes (e.g., background) as a false positive in mIoU computation. In comparison, our revised mIoU includes these pixels to avoid model cheating by predicting more ignored classes. The experiments were conducted on NYUv2 [42], ScanNet [12], indoor ADE20k [54], and indoor SUNRGB-D [43] datasets. We filtered ADE20k to include only indoor scenes and ignored previously omitted classes in an NYUv2 setting following the prior work [25] for SUNRGB-D. More details on our split are provided in Appendix D.

When pre-trained on non-paired RGB datasets, Ours (pseudo) outperforms all 2D competitors in both architectures, including DINO [3] and MAE [20]. For 2D-3D competitors, we win against Pri3D and achieve very competitive results to Set-InfoNCE in NYUv2 by only -0.16 in mIoU [25]. We also outperform Pri3D in ADE20k by +1.01 for the original and +0.79 for the pseudo version. The qualitative results are shown in Fig. 6 and Appendix E.

For ScanNet and SUNRGB-D datasets, Pri3D and Set-InfoNCE perform better than ours. This might be because their models were trained directly on ScanNet 3D scans. Some images in SUNRGB-D also appear similar to those in ScanNet, as both datasets sampled RGB-D frames from recorded videos at a high frame-per-second rate. Nonetheless, even without 3D ScanNet, our method still achieves comparable results to Pri3D (-0.17).

Additionally, in Appendix F, we present results obtained by increasing the pre-training data, which allows us to surpass Set-InfoNCE on NYUv2 without real 3D pairs.

Table 1. **Semantic segmentation results on four popular benchmark datasets.** For both ResNet-50 (RN50) and ViT-B architectures, our approaches have better mIoU than 2D representation learning baselines. (*) denotes scores from the methods that were directly pre-trained on 2D and 3D data of ScanNet dataset. We win against Pri3D in NYUv2 and ADE20k without requiring 3D scene scans.

				NYUv2		**ScanNet**		**indoor ADE20k**		**SUNRGB-D**	
Arch.	**GT pair**	**Method**	**3D**	**mIoU**	**mIoU [25]**	**mIoU**	**mIoU [25]**	**mIoU**	**mIoU [25]**	**mIoU**	**mIoU [25]**
RN50	2D only	SupImg	-	47.04	50.0	49.12	55.7	37.52	39.69	47.96	63.12
		SimCLR	-	47.94	53.32	49.91	56.34	38.19	39.56	48.51	63.00
		SupCon	-	48.49	54.16	50.59	56.57	37.96	40.13	48.13	65.01
		SupCon (fine)	-	47.34	53.45	49.81	56.22	37.87	39.92	48.05	63.13
	pseudo	*Ours (pseudo)*	CAD	49.46	54.62	50.77	56.76	39.13	40.43	49.13	65.45
	2D-3D	CrossPoint	CAD	46.04	49.18	47.26	54.02	36.55	37.88	47.15	62.77
		Pri3D	scene	49.52	54.7	54.72*	61.7*	38.34	39.17	**50.02**	**66.65**
		Set-InfoNCE	scene	-	**55.4**	-	**63.1***	-	-	-	-
		Ours	CAD	**49.77**	55.24	51.03	57.12	**39.35**	**40.86**	49.85	65.95
ViT-B	2D only	SupImg	-	49.44	55.59	63.93	59.72	42.78	44.38	51.26	66.45
		DINO	-	52.14	57.24	62.54	58.13	41.72	43.41	50.52	66.12
		MAE	-	50.10	55.66	62.23	58.95	42.96	44.96	52.22	67.87
	pseudo	*Ours (pseudo)*	CAD	52.47	58.02	64.49	60.73	43.02	45.37	52.53	69.59
	2D-3D	*Ours*	CAD	53.0	58.67	65.27	60.95	43.12	45.82	52.96	69.69

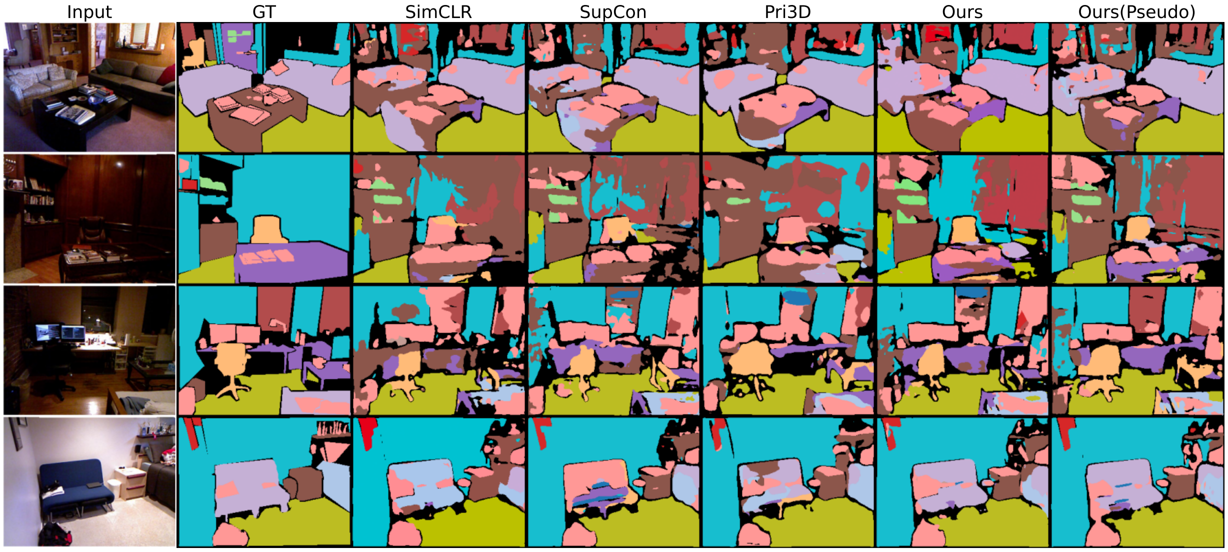

Figure 6. **Qualitative results on NYUv2 [42] semantic segmentation.** Our methods yield better segmentation results when labeling a scene is challenging due to color or lighting. The geometric priors in our learned representations boost understanding of each scene.

5.3. 2D Instance segmentation and object detection

Indoor scenes. We show how our features improve instance segmentation and object detection results in Table 2. Similar to [25], we use Mask-RCNN [22] with ResNet-50 encoder, implemented on Detectron2 [48] framework for both tasks. For ViT-B, we resize input images to 224 × 224 and follow [33] by using Mask-RCNN as the detection head. All backbones were initialized by the pre-trained weights of our model or competitors'. Following [5, 25], we use NYUv2 [42] dataset and a well-known detection benchmark, COCO dataset, with Average Precision (AP) as the metric. We filtered COCO images to include only indoor objects. Details on the selected classes are in Appendix D.

Our method outperforms state-of-the-art 2D and 2D-3D competitors in all settings. Ours (pseudo) achieves competitive scores against Pri3D and Set-InfoNCE in NYUv2 and beats Pri3D in COCO without ground truth 3D pairs.

Outdoor scenes. We also pre-trained our model on PASCAL3D+ [50] dataset, which contains objects such as vehicles and furniture in both indoor and outdoor scenes, and evaluated the model on another subset of COCO images called outdoor COCO. Further dataset details are provided in Appendix D. Our method leads to similar improvements in AP as observed in the indoor-only settings.

5.4. Experiments on 2D retrieval

Table 3 shows the retrieval performance on Pix3D [45] by fine-tuning the pre-trained model on coarse-grained (category) and fine-grained (subcategory) labels of objects. The 2D encoder was fine-tuned along with an additional retrieval head, consisting of two linear layers of sizes 1024 and 512 with ReLU applied only to the first layer. All

Table 2. **Object detection and instance segmentation performance.** We outperform competitors for all datasets.

					NYUv2						indoor COCO						outdoor COCO					
					Object Det.			**Instance segm.**			**Object Det.**			**Instance seg.**			**Object Det.**			**Instance seg.**		
Arch.	**Size**	**GT pair**	**Method**	**3D**	**AP50**	**AP75**	**AP**	**AP50**	**AP75**	**AP**	**AP50**	**AP75**	**AP**	**AP50**	**AP75**	**AP**	**AP50**	**AP75**	**AP**	**AP50**	**AP75**	**AP**
RN50	480	2D only	SupImg	-	29.9	17.3	16.8	25.1	13.9	13.4	41.78	24.21	23.70	39.16	23.35	22.61	46.09	26.98	28.08	42.45	23.34	23.92
			SimCLR	-	32.81	20.15	19.24	29.10	15.97	15.62	43.63	26.46	25.45	40.87	24.79	23.86	48.15	28.75	30.40	44.31	25.01	24.99
			SupCon	-	33.23	20.36	19.63	29.44	16.16	15.83	43.66	26.34	25.32	40.84	24.53	23.75	47.89	28.67	30.29	44.16	24.63	24.97
			SupCon (fine)	-	32.56	19.74	18.92	29.06	16.11	15.74	43.58	25.95	25.21	40.65	24.22	23.66	45.01	27.90	26.59	41.97	25.61	24.66
		pseudo	*Ours (pseudo)*	CAD	34.45	20.27	19.72	29.64	16.24	16.13	43.74	26.47	25.48	40.92	24.77	23.91	-	-	-	-	-	-
		2D-3D	CrossPoint	CAD	28.42	15.94	15.22	24.49	13.32	13.11	40.25	22.78	22.26	38.54	21.92	20.80	43.22	24.57	25.60	39.75	21.93	21.11
			Pri3D	scene	34.0	20.4	19.4	29.5	16.3	15.8	43.49	26.40	25.22	40.71	24.72	23.61	-	-	-	-	-	-
			Set-InfoNCE	scene	34.6	20.5	19.7	29.7	16.3	16.5	-	-	-	-	-	-	-	-	-	-	-	-
			Ours	CAD	**34.85**	**20.89**	**20.12**	**30.03**	**16.51**	**16.84**	**44.11**	**26.78**	**25.69**	**41.02**	**24.91**	**24.08**	**49.03**	**29.80**	**31.62**	**45.23**	**25.90**	**25.85**
ViT-B	224	2D only	SupImg	-	34.40	19.24	19.06	28.42	14.05	14.97	31.45	20.63	19.41	29.77	18.73	17.82	33.56	23.19	21.81	31.68	19.52	18.11
			DINO	-	33.03	18.62	17.91	26.82	14.56	14.73	27.70	16.24	15.87	25.78	14.86	14.76	32.57	22.13	20.61	29.86	18.07	17.66
			MAE	-	35.92	19.30	19.24	29.88	16.01	15.82	31.54	20.59	19.33	29.92	18.65	17.83	36.97	24.51	23.12	33.67	20.15	19.46
		pseudo	*Ours (pseudo)*	CAD	36.24	19.78	19.72	30.10	15.94	16.05	31.78	20.74	19.46	30.01	19.07	17.94	-	-	-	-	-	-
		2D-3D	*Ours*	CAD	36.31	19.91	19.94	30.30	16.16	16.27	32.02	21.04	19.67	30.16	19.02	18.09	37.74	24.92	23.42	34.13	20.49	19.89

Table 3. **Retrieval performance on Pix3D [45].** Our method outperforms all competitors in all situations.

Arch.	**GT Pair**	**Method**	**3D**	**Coarse R@1**	**Fine R@1**
RN50	2D only	SupImg	-	78.85	51.18
		SimCLR	-	80.04	55.01
		SupCon	-	81.32	52.39
		SupCon (fine)	-	79.95	55.84
	pseudo	*Ours (pseudo)*	CAD	81.74	56.71
	2D-3D	CrossPoint	CAD	75.97	48.09
		Pri3D	scene	79.72	51.63
		Ours	CAD	**82.88**	**58.97**
ViT-B	2D only	SupImg	-	84.04	64.17
		DINO	-	84.59	65.48
		MAE	-	83.79	62.00
	pseudo	*Ours (pseudo)*	CAD	86.16	66.40
	2D-3D	*Ours*	CAD	86.32	67.23

Table 4. **Effects on losses and positive mining choices.** Training on three losses simultaneously yields the best performance in mIoU. At the same time, choosing the Chamfer Distance (CD) as a positive mining choice in L_{GEO} also wins against the others.

L_{IMG}	L_{GEO}	L_{CROSS}	Pos.	**NYUv2**	**ADE20k**
-	✓	✓	CD	48.23	37.71
✓	-	✓	-	48.46	38.47
✓	-	-	-	47.94	38.19
-	-	✓	-	47.02	37.43
✓	✓	✓	Sup	49.48	39.12
✓	✓	✓	Aug	49.14	38.65
✓	✓	✓	CD	**49.77**	**39.35**

retrieval models were trained using supervised contrastive loss [28] with $\tau = 0.07$. Our methods outperform all competitors in all architectures.

5.5. Ablation studies

All ablation experiments are evaluated on NYUv2 and ADE20k semantic segmentation tasks using our revised mIoU metric on ResNet-50 architecture.

Effects on losses and positive mining choices Table 4 shows the effects of three loss choices (i.e., L_{IMG}, L_{GEO}, and L_{CROSS}) and three different choices of the positive point cloud using in Eq. (2), including using object categories (Sup), point cloud augmentation (Aug), and Chamfer distance (CD). We found that training three loss functions simultaneously notably improves mIoU from training single or dual losses. While in the positive choice study, our selected Chamfer distance achieves the best mIoU.

Effects on K selected positive point clouds. We studied the best number of K in Eq. (2). Fig. 7 reveals that $K = 3$ can achieve the best mIoU among the others. Using a lower K leads to lower mIoU, similar to a higher K. This is probably because higher K has more chance to select false positive samples due to higher Chamfer distances, which means that the selected samples might have a much different shape than the anchor $\mathbf{z}_i^G$.

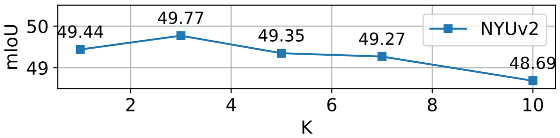

Figure 7. **Effect of varying K selected positive point clouds.** $K = 3$ has the best mIoU.

6. Limitations and Discussion

Using CAD models for inducing geometric-aware properties in 2D representation via Chamfer distance shows significantly enhanced 2D object understanding results that are almost as effective as 3D priors from scene scans. This finding remains true when we train our model on generated pseudo-RGB-CAD pairs, allowing us to beat SOTA 2D competitors without the need for paired datasets. However, the performance of the pseudo-pair generator currently hinders our results. As seen in Fig. 2 and 4, features trained on pseudo pairs cannot clearly categorize intra-subcategory images. We observe that this is due to an improper CAD pair assignment; more information is given in Appendix B. Nonetheless, generating or retrieving 3D models from a single image is still more tractable than scene scans and shows progressive performance improvement [38]. This enables harnessing the massive-scale RGB data and generalizing beyond domains with paired data (See Appendix F).

References

[1] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9902–9912, June 2022. 2, 3, 5

[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

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 1, 2, 5, 6

[4] 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

[5] Nenglun Chen, Lei Chu, Hao Pan, Yan Lu, and Wenping Wang. Self-supervised image representation learning with geometric set consistency. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR*, 2022. 1, 3, 5, 6, 7

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, 2020. 1, 2, 3, 4, 5

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 2

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020. 1

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 1, 2

[11] Ye Chen, Jinxian Liu, Bingbing Ni, Hang Wang, Jiancheng Yang, Ning Liu, Teng Li, and Qi Tian. Shape self-correction for unsupervised point cloud understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8382–8391, 2021. 2

[12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 3, 5, 6

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2009. 1, 5, 6

[14] Benjamin Eckart, Wentao Yuan, Chao Liu, and Jan Kautz. Self-supervised learning on 3d point clouds by learning discrete generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8257, 2021. 2

[15] Yutong Feng, Yifan Feng, Haoxuan You, Xibin Zhao, and Yue Gao. Meshnet: Mesh neural network for 3d shape representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8279–8286, 2019. 2

[16] Georgia Gkioxari, Justin Johnson, and Jitendra Malik. Mesh r-cnn. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 2, 3, 5

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1, 2, 4

[18] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. A papier-mache approach to learning 3d surface generation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 216–224, 2018. 3

[19] Can Gümeli, Angela Dai, and Matthias Nießner. Roca: Robust cad model retrieval and alignment from a single image. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2022. 2, 3, 5

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 5, 6

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. 1, 2

[22] 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. 5, 7

[23] 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. 2, 5

[24] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597, 2021. 2

[25] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Niesner. Pri3d: Can 3d priors help 2d representation learning? *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021. 1, 2, 3, 5, 6, 7

[26] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal center loss for 3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3142–3151, 2021. 3

[27] Longlong Jing, Ling Zhang, and Yingli Tian. Self-supervised feature learning by cross-modality and cross-view correspondences. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1581–1891, 2021. 2, 3

[28] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 2020. 1, 2, 3, 5, 8

[29] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 5

[30] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Mask2cad: 3d shape prediction by learning to segment and retrieve. In *European Conference on Computer Vision*, pages 260–277. Springer, 2020. 3, 5

[31] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Patch2cad: Patchwise embedding learning for in-the-wild shape retrieval from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12589–12599, 2021. 2, 3, 5

[32] Florian Langer, Ignas Budvytis, and Roberto Cipolla. Leveraging geometry for shape estimation from a single rgb image. In *BMVC*, 2021. 3

[33] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *ECCV*, 2022. 7

[34] Ming-Xian Lin, Jie Yang, He Wang, Yu-Kun Lai, Rongfei Jia, Binqiang Zhao, and Lin Gao. Single image 3d shape retrieval via cross-modal instance and category contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11405–11415, 2021. 3

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[36] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. 3d-to-2d distillation for indoor scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4464–4474, 2021. 3

[37] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, 2022. 2

[38] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 8

[39] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6

[42] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 1, 2, 5, 6, 7

[43] 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. 6

[44] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 2

[45] 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, 7, 8

[46] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 38(5):1–12, Nov 2019. 2, 4, 5

[47] Xin Wen, Junsheng Zhou, Yu-Shen Liu, Zhen Dong, and Zhizhong Han. 3d shape reconstruction from 2d images with disentangled attribute flow. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3793–3803, 2022. 3, 5

[48] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 7

[49] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[50] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014. 7

[51] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understand-

ing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 6

[52] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European conference on computer vision*, pages 574–591. Springer, 2020. 2

[53] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021. 2

[54] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2, 6