Appendix

In this appendix, we provide additional clarifications and experiments as follows:

- Section A: Additional training details
- Section B: Additional studies on geometric-aware properties in the learned representation
- Section C: Additional details on multiple object localization in the learned representation
- Section D: Datasets used in the downstream tasks
- Section E: Semantic segmentation results
- Section F: Experiments on larger pre-train datasets
- Section G: Effects of the choice of pseudo pair selection policy.
- Section H: Effects of the choice of CAD model preprocessing method
- Section I: Potential negative societal impacts

A. Additional training details

In the ResNet-50 [11] backbone setting, we trained our proposed model using an SGD optimizer with a learning rate of 0.001 and a momentum of 0.9, decayed by a polynomial decay scheduler. In the ViT-B [15] setting, we trained our model using AdamW [18] optimizer with a learning rate of 0.0001, $\beta_1 = 0.9$, and $\beta_2 = 0.99$ using a polynomial decay scheduler.

B. Additional details on geometric-aware properties in the learned representation

B.1. Geometric-aware properties in learned representations of other categories

Fig. 1 extends Fig. 2 in the main paper to include more object categories in addition to the Sofa class. This Fig. 1 visualizes the pairwise cosine similarities between the encodings of objects from several categories in the Pix3D [21] validation set.

B.2. Geometric-aware properties in learned representations of other competitors

In Fig. 2, we provide an additional visualization of the pairwise cosine similarities from our method and more competitors in the ResNet-50 2D backbone setting. The competitors are alternative positive point cloud choices in L_{GEO} (Eq. 2) using point cloud augmentations (i.e., Aug) and supervised discrete signals like object labels (i.e., Sup), which are the same baselines used in the ablation study in Section 5.5.1 of the main paper.

We found that Aug achieves geometric awareness but performs worse than ours in inter and intra-category pairwise similarities, as shown in Fig. 4 in the main paper. On the other hand, Sup fails to group intra-subcategories images. Such drawbacks may cause these alternatives to perform poorly during fine-tuning and on the downstream tasks in the main paper.

Fig. 2 also visualizes the learned representations from our model using ViT-B 2D encoder backbone. The encoded features from ViT-B show better main category and subcategory grouping and discrimination than ResNet-50. Compared to the recent state-of-the-art 2D representation learning, DINO [1] and MAE [10], both Ours and Ours (pseudo) clearly show better geometric-aware properties.

B.3. Geometric-aware properties in learned representations of a model trained on pseudo pairs

We further demonstrate the effectiveness of learning representation on pseudo-RGB-CAD pairs over off-the-shelf self-supervised representation learning.

We first start by analyzing the performance of our model trained on pseudo pairs (i.e., Ours (pseudo)). As in Fig. 1 and Fig. 2 of the main paper, Ours (pseudo) struggles with grouping intra-subcategory images and discriminating between inter-subcategory images. We observe that our pseudo-pair generator, ROCA [9], retrieves CAD models with limited model variations and has moderate retrieval accuracy (54-87%). Only 64.33% of Pix3D [21] images were correctly paired with CAD models in the same category. We suspect that these poor assignments can be because ROCA uses a different CAD database, i.e., ShapeNet [2], which is different from those used in Pix3D, which are IKEA products. These incorrect assignments directly affect positive pairing used in the training objectives, leading to weaker geometric awareness compared to that obtained from real RGB-CAD pairs.

We then conduct additional experiments to evaluate whether Ours (pseudo), which is trained with imperfect CAD assignments, can *really* help induce geometric-aware properties in the 2D representation. Table 2 shows the mean pairwise cosine similarity of objects within the same subcategory and across subcategories within the same category, averaged across all 395 subcategories in the Pix3D validation set.

We found that Ours (pseudo) can improve the mean cosine similarity from the baseline with no 3D priors, SimCLR [4] using the ResNet-50 backbone. Compared to SimCLR, our representations have a 1.71% lower inter-subcategory mean and a 2.34% higher difference between the means of inter-subcategory and intra-subcategory. We achieve larger inter-intra differences in 241 out of 395 subcategories, which indicates that our representations can better discriminate one subcategory from other subcategories. We visualize the distribution of pairwise cosine similarities in Fig. 3. Such improvements in intra-subcategory differences and intra-inter-differences reflect improved geometric awareness and could help improve 2D object understanding

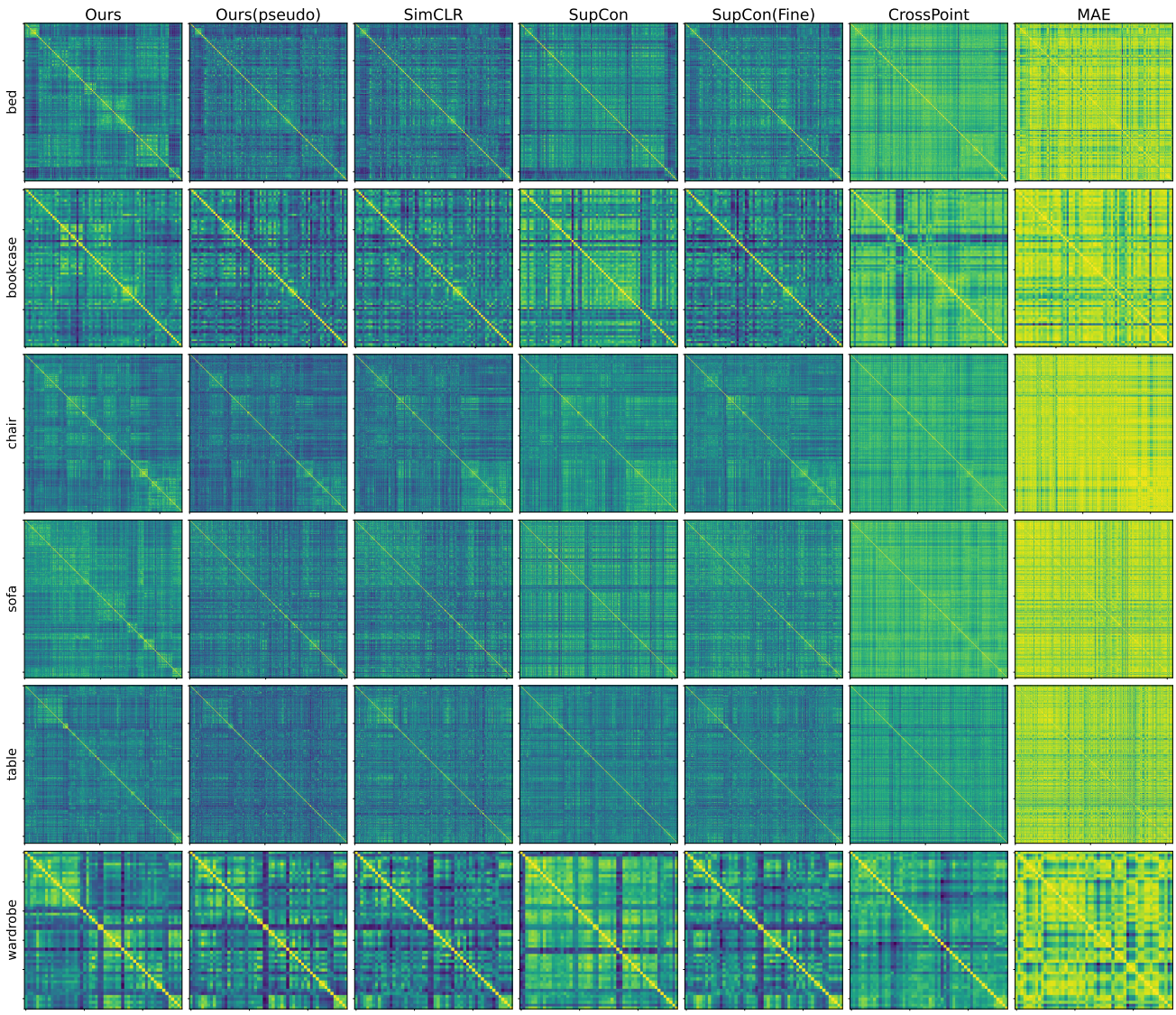


Figure 1. Visualization of the pairwise cosine similarities between the learned representations of objects from each method. The bright color indicates higher similarity. Each row shows the similarities of validation images in Pix3D [21] for each category, sorted by object's subcategory (i.e., object models). Our method shows a better grouping of the same subcategory than others for all category types.

results, as demonstrated by the experiments in the main paper.

C. Additional details on multiple object localization in the learned representation

In Section 5.1 of the main paper, we demonstrated how our method handles images with multiple objects. To achieve this, we use encoded features extracted from the ViT-B backbone to create a codebook for object category prediction and patch-wise features of unseen inference images.

We construct the codebook by extracting the global feature (i.e., [CLS] embedding) of each Pix3D image from the ViT-B encoder. We then compute the mean [CLS] embedding of each category by averaging embeddings of all images belonging to the same category. The codebook contains the means for seven object categories in Pix3D, including bed, bookcase, chair, desk, sofa, table, and wardrobe.

For the inference image, we randomly select one from the NYUv2 dataset and encode it using the ViT-B encoder to obtain a patch-wise feature with a size of 196×768.

Given the codebook and the inference image, we compute the cosine similarity between each patch-wise feature from the image and each mean category feature in the codebook. This enables us to obtain the object category with the

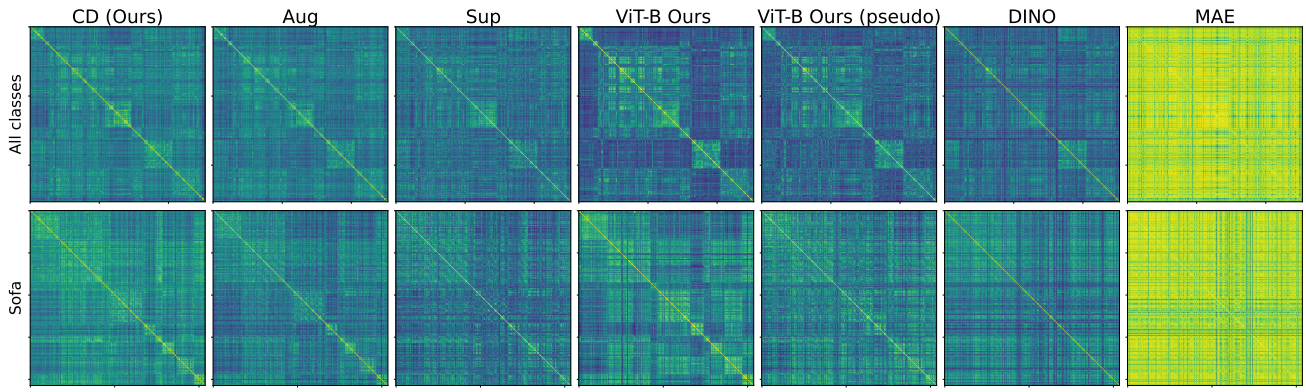


Figure 2. Visualization of the pairwise cosine similarities between the learned representations of objects from each method. The bright color indicates higher similarity. The first row shows the similarities of all validation images in Pix3D [21] dataset, sorted by object category and subcategory (i.e., object models). The second row zooms in on a single category (Sofa) from the first row. Point cloud augmentation (Aug) and supervised category labels (Sup) are less capable of grouping the same sofa than Ours, which are trained using geometric priors based on the Chamfer distance. While in the ViT-B [15] backbone setting, our methods also clearly show better grouping between the same sofa and better category classification than SOTAs like DINO [1] and MAE [10].

Classes	Original CADs #	Predictied CADs #	Matching acc.
All pred. classes	351	42	64.33
bed	20	6	53.76
bookcase	17	2	86.97
chair	221	3	61.68
sofa	20	19	83.09
table	63	2	72.38
wardrobe	10	10	84.38

Table 1. **Performance of pseudo-pair generator (ROCA [9]) on Pix3D [21] dataset.** ROCA predicts CAD models with very limited model variations for some categories compared to the original ground-truth variations. The matching accuracy is also low for some categories.

Metrics	SimCLR	*Ours (pseudo)*
Intra-subcategory mean ↑	0.8057 (172)	**0.8121 (223)**
Inter-subcategory mean ↓	0.58.97 (115)	**0.5798 (280)**
Intra-Inter differences ↑	0.2223 (154)	**0.2275 (241)**

Table 2. **The mean pairwise cosine similarity of the learned representations for each of the 395 subcategories in the Pix3D dataset.** Intra-Inter differences denote the differences between the mean pairwise cosine similarity of intra-subcategory images and inter-subcategory images of each subcategory. The numbers in the parentheses show the numbers of subcategories with better (higher or lower) values.

highest similarity score for each patch. We then use these predicted patch-wise categories to create the segmentation map visualization shown in Fig. 6 of the main paper.

Additional results from different NYUv2 inference images are presented in Fig. 4

D. Datasets used in the downstream tasks

D.1. NYUv2 [19]

The NYUv2 dataset provides recorded RGB-D frames of indoor scenes and labeled segmentation masks, instance segmentation masks, and object detection bounding boxes. For the evaluations in the main paper, we followed the procedure in prior works [3, 12] and used the official train-test split (795 for training and 654 for testing).

D.2. ScanNet [5]

The ScanNet dataset provides 3D scans of indoor scenes, their corresponding recorded RGB-D videos, and ground truths, e.g., segmentation masks for various 3D and 2D scene understanding tasks. Similar to prior works [3, 12], we used the standard ScanNet 2D semantic segmentation benchmark that samples 25,000 RGB frames (20,000 for training and 5,000 for testing) from the RGB-D videos for evaluating semantic segmentation fine-tuning performance.

D.3. ADE20k [23]

The ADE20k (2016) dataset consists of RGB scene images from various places and their semantic segmentation labels. In our experiments, we filtered the dataset to contain only indoor scenes for a fair comparison with two pretraining datasets used in the experiments (i.e., Pix3D [21] and ScanNet [5]). In particular, we selected 9 scene categories to include: attic, bedroom, child room, dining room, dorm room, closet, hotel room, living room, and television room. This selection results in 2951 training images and

296 validation images used in our experiments. We also ignored classes that were never seen in the pre-trained models, resulting in eight classes: bed, cabinet, table, chair, sofa, desk, wardrobe, and bookcase.

D.4. SUNRGB-D [20]

The SUNRGB-D dataset consists of 10,000 RGB-D frames (5,000 for training and 5,000 for testing) with segmentation mask labels. We excluded a total of 24 classes since SUNRGB-D segmentation masks contain 38 classes of segmentation labels with several categories that were not provided in our pre-trained dataset (i.e., Pix3D). This ends up in 14 classes using in our fine-tuning experiments: wall, floor, cabinet, bed, chair, sofa, table, door, window, bookshelf, picture, desk, ceiling, and background. These classes are the same as the NYUv2 class filtering setting provided in [12].

D.5. COCO [17]

COCO [17] is a well-known object detection and instance segmentation benchmark with various types of objects (e.g., objects in indoor and outdoor scenes, sports equipment, kitchenware, person, or animals).

Indoor COCO. In our experiments, we use COCO-2014 split, including only images with 24 object classes related to indoor scene objects. The chosen classes consist of chairs, beds, TV, remotes, microwave, sink, clocks, teddy bears, couches, dining tables, laptops, keyboards, ovens, refrigerators, vases, hair drier, potted plants, toilets, mice, cell phones, toaster, book, scissors, and toothbrush. This split contains 32,186 training images and 15,954 validation images.

Outdoor COCO. We use the same COCO-2014 train-test split and selected a different set of images consisting of 12 object classes that are associated with both indoor and outdoor scenes. These classes include bicycle, airplane, truck, couch, car, bus, boat, dining table, motorcycle, train, chair, and TV. This split contains 82,783 training images and 40,504 validation images.

D.6. PASCAL3D+ [22]

The PASCAL3D+ dataset consists of RGB-CAD pairs of 12 categories that mix indoor and outdoor objects, including airplane, bicycle, boat, bottle, bus, car, chair, dining table, motorbike, sofa, train, and TV. The official train-test split was used for pre-training, with 22,054 training images and 7,352 validation images. The RGB images in this dataset were sourced from ImageNet [6] and PASCAL VOC 2012 [7] dataset.

D.7. Pix3D [21]

Pix3D was used in both the pre-training stage and the fine-tuning stage in the object retrieval task in the main paper. Pix3D consists of 10,069 images of indoor objects categorized into eight main categories (i.e., bed, bookcase, chair, desk, sofa, table, tool, wardrobe, miscellaneous) and 395 subcategories based on furniture models. Each image represents one primary object with its associated CAD model. In object retrieval fine-tuning, we used S1 train-test split [8] with 7,539 training images and 2,530 validation images. We trained our model using their given main category or subcategory labels, then tested the model using Coarse Recall@1 for the main category or Fine Recall@1 for the subcategory.

E. Semantic segmentation results

We provide additional qualitative segmentation results on NYUv2, ADE20k, SUNRGB-D, and ScanNet datasets in Fig. 5, Fig. 6, Fig. 7, and Fig. 8, respectively. All of the results are from the ResNet-50 backbone setting. Our methods show better segmentation results than 2D representation learning baselines. In some cases, we also outperform Pri3D [12], which requires 3D scene scans.

F. Experiments on larger pre-train datasets

We additionally pre-trained our model on an expanded dataset by generating pseudo pairs for each indoor image in the ImageNet [6] and COCO [17] datasets. This resulted in a larger RGB-CAD dataset than the original training set (i.e., Pix3D), which contains only 7,359 images.

ImageNet [6]. We selected 11,875 images from 11 classes (bookcase, bathtub, studio couch, file, china cabinet, folding chair, rocking chair, barber chair, dining table, and monitor) in the official training split of ImageNet. The selected classes are chosen to be matched with Pix3D's classes.

COCO [17]. We use the same protocol in D.5 to filter the dataset. We chose only 13 classes (chairs, beds, clocks, teddy bears, toilets, couches, dining tables, vases, hair drier, potted plants, books, scissors, and toothbrushes). The chosen classes consist of 26,396 images. Like the setting in ImageNet, these selected classes are also resembling Pix3D's classes.

By combining these two datasets with Pix3D, we obtained a total of 44,903 pairs, which is 5.95 times larger than Pix3D.

Table 3 compares the fine-tuning results of our method in ResNet-50 backbone trained on the original Pix3D dataset and this new dataset, referred to as Mixed. Training our method on the Mixed dataset, which contained more available RGB-CAD pairs, can improve semantic segmentation performance on NYUv2 compared to training on the original dataset.

Pre-train dataset	GT pair	Method	NYUv2 mIoU	NYUv2 mIoU [12]
Pix3D	2D only	SimCLR	47.94	53.32
	pseudo	*Ours (pseudo)*	49.46	54.62
	2D-3D	*Ours*	49.77	55.24
ScanNet	2D-3D	Pri3D	49.52	54.7
		Set-InfoNCE	-	55.4
Mixed	2D only	SimCLR	48.97	54.16
	pseudo	*Ours (pseudo)*	**49.92**	**55.55**

Table 3. **Semantic segmentation results of ours and competitors trained on different pre-train datasets.** Training our method on a larger RGB-CAD dataset (Mixed) yields the best results.

Choices	NYUv2 mIoU	ADE20k mIoU
Cropping	49.06	38.56
Highest conf.	48.44	37.95
Largest bbox	**49.46**	**39.13**

Table 4. **Effects of predicted pseudo-pair choices.** Selecting the largest bounding box yields the best performance.

G. Effects of the choice of pseudo pair selection policy

We provide an ablation study that shows the effect of changing the method to construct each pseudo-RGB-CAD pair in ROCA [9]. After detecting a set of objects from a given RGB image using Mask-RCNN, we have to select one (or more) objects from the prediction to represent the associated CAD model of the given image. We define three policies for selecting the object:

Cropping all detected boxes: We use all detected objects in the prediction to construct multiple RGB-CAD pairs. The input RGB image is cropped based on the predicted bounding box of each detected object and then further paired with its predicted CAD model. This results in multiple RGB-CAD pairs for each input RGB image.

Highest confidence: This policy selects the CAD model with the highest confidence score given by Mask-RCNN. This results in one RGB-CAD pair for each input RGB image.

Largest detected box: This policy selects the CAD model with the largest bounding box. This also results in one RGB-CAD pair for each input.

Table 4 shows the experimental results evaluated on the NYUv2 and indoor ADE20k semantic segmentation tasks using our revised mIoU metric on ResNet-50 architecture. We found that using the largest detected box policy yields the best performance.

Nat. Img.	Rend. Img.	PC	NYUv2 mIoU	ADE20k mIoU
✓	-	-	47.94	38.19
✓	✓	-	47.53	38.22
✓	-	✓	**48.46**	**38.47**
✓	✓	✓	48.09	38.36

Table 5. **Effects of CAD 3D data structure choices.** Cross-learning between natural images (Nat. Img.) and point clouds (PC) yields the best performance.

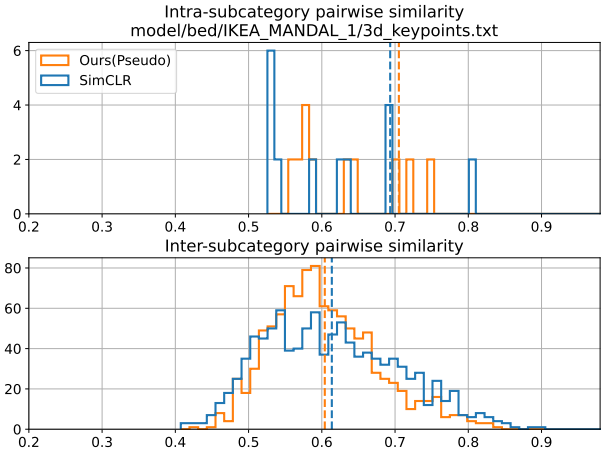
H. Effects of the choice of CAD model preprocessing method

This experiment evaluates an alternative CAD model preprocessing method that renders CAD models into 2D images used in prior works [13, 14, 16] against point clouds used in our work. We also show whether training on RGB images with both options at the same time (i.e., trimodal learning) can further enhance downstream performance.

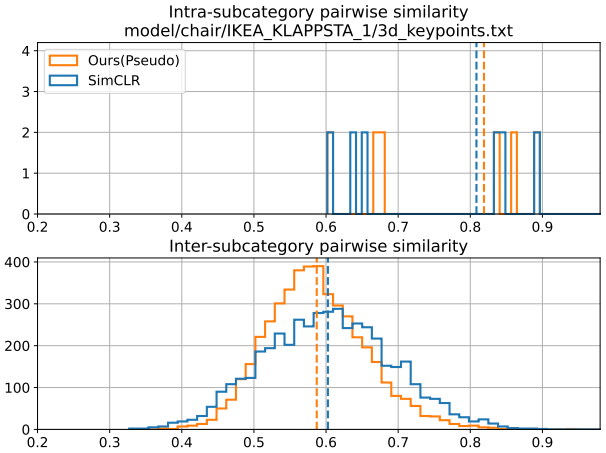
For rendered images, we use a 2D encoder based on ResNet-18 [11] and a modality-specific projection head to encode four rendered views of each input, similar to [13]. The fine-tuning results shown in Table 5 are evaluated on the NYUv2 and indoor ADE20k semantic segmentation tasks using our revised mIoU metric on ResNet-50 architecture. We found that training on RGB images along with point clouds achieves better results than using rendered images. In addition, using natural images, rendered images, and point clouds all together has shown decreased task performance.

I. Potential negative societal impacts

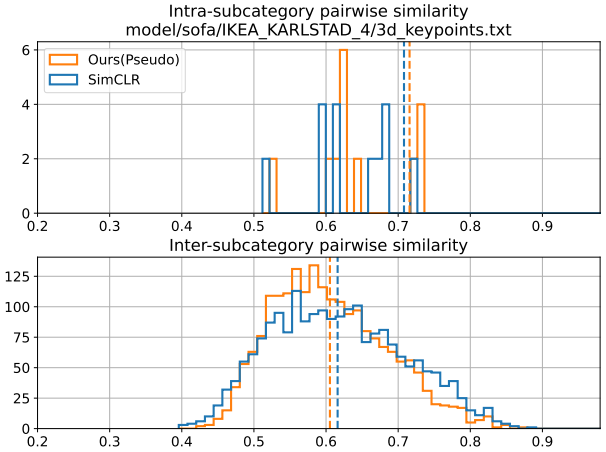
Similar to 2D object detection or recognition algorithms trained on predefined object categories, our models, once fine-tuned on downstream tasks, may not generalize to unseen categories outside the training set. This could lead to social exclusion. For example, the models might ignore specific items from minority groups, countries, or cultures, leading to cultural discrimination. Nonetheless, our method's ability to learn from generated RGB-CAD pairs enables training on a wide variety of domains where paired datasets do not exist. This can help scale up the number of known classes and increase social inclusion.



(a) Subcategory pairwise similarities of IKEA MANDAL 1 (bed)



(b) Subcategory pairwise similarities of IKEA KLAPPSTA 1 (chair)



(c) Subcategory pairwise similarities of IKEA KARLSTAD 4 (sofa)

Figure 3. **Distribution of pairwise cosine similarities among intra-subcategory and inter-subcategory samples.** The higher mean intra-subcategory similarity and the lower mean inter-subcategory similarity in Ours (pseudo) demonstrate better discrimination between each subcategory compared to SimCLR.

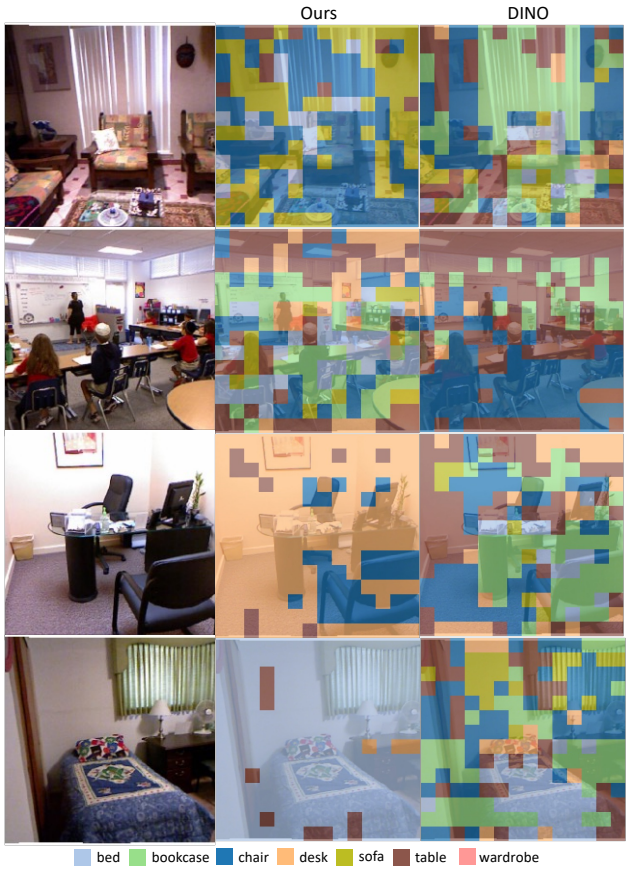


Figure 4. More qualitative results of unsupervised multiple object localization on NYUv2 [19] dataset.

Figure 5. More qualitative results of semantic segmentation on NYUv2 [19] dataset.

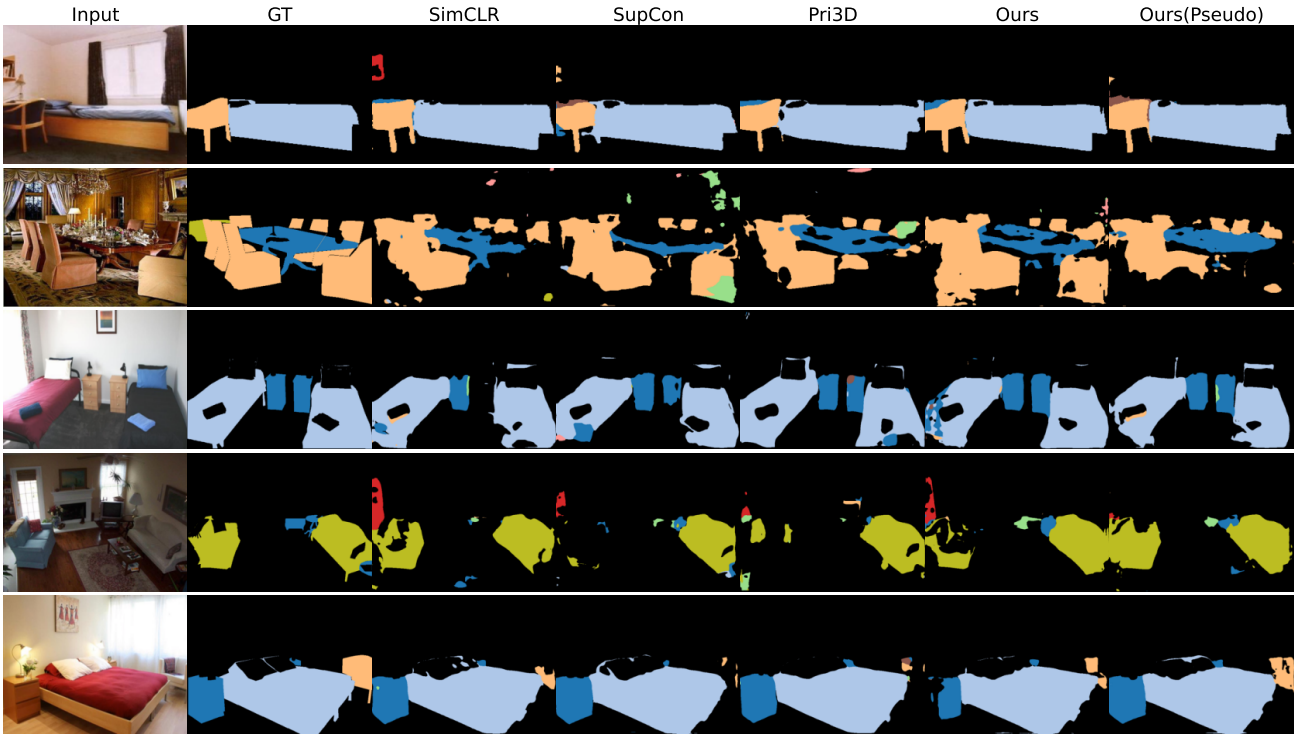


Figure 6. Qualitative results of semantic segmentation on ADE20k [23] dataset.

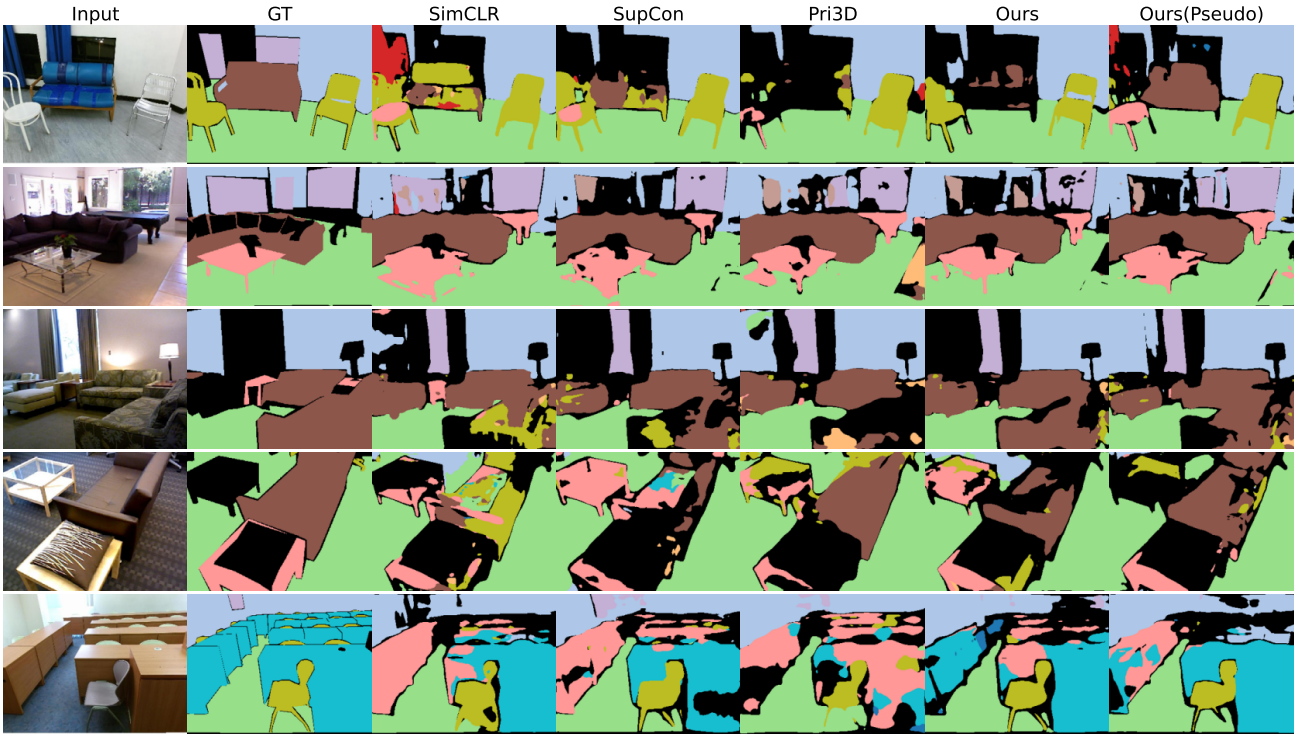


Figure 7. Qualitative results of semantic segmentation on SUNRGB-D [20] dataset.

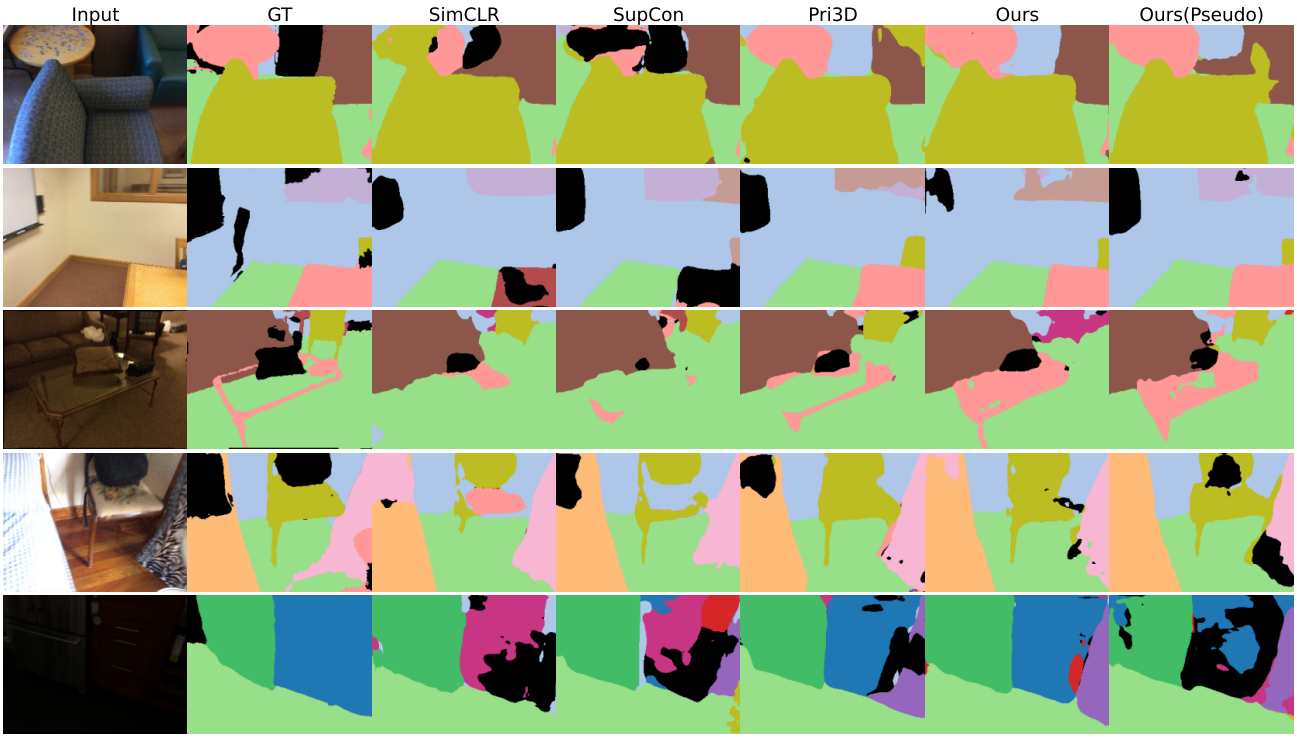


Figure 8. Qualitative results of semantic segmentation on ScanNet [5] dataset.

References

[1] 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, 3

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[3] 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. 3

[4] 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

[5] 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. 3, 8

[6] 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. 4

[7] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, jun 2010. 4

[8] 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. 4

[9] 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. 1, 3, 5

[10] 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, 3

[11] 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, 5

[12] 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. 3, 4, 5

[13] 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. 5

[14] 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. 5

[15] 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. 1, 3

[16] 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. 5

[17] 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. 4

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 1

[19] 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. 3, 6, 7

[20] 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. 4, 8

[21] 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. 1, 2, 3, 4

[22] 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. 4

[23] 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. 3, 7