Pri3D: Can 3D Priors Help 2D Representation Learning?
Supplementary Material

Ji Hou[1] Saining Xie[2] Benjamin Graham[2] Angela Dai[1] Matthias Nießner[1]
[1]Technical University of Munich [2]Facebook AI Research

In this appendix, we show data-efficient learning results on more downstream tasks spanning different percentages of used training images in Section 1. We show more visualizations of semantic segmentation results in Section 2. To provide more results across datasets, we further demonstrate a variety of experimental results of outdoor data in Section 3. As promised in main paper, we show the results of **Unsupervised Pri3D** in Section 4. In the end, we discuss the current limitations and potential directions of improvements in Section 5.

1. Data-Efficient Learning

We plot the curves across different percentages of used training data for data-efficient learning similarly to [8]. We demonstrate that our pre-training algorithm generalizes with different backbones. We show data-efficient learning curves on semantic segmentation task in ScanNet [3] with a ResNet18 [7] backbone in Figure 1. Please refer to main paper for results with a ResNet50 backbone.

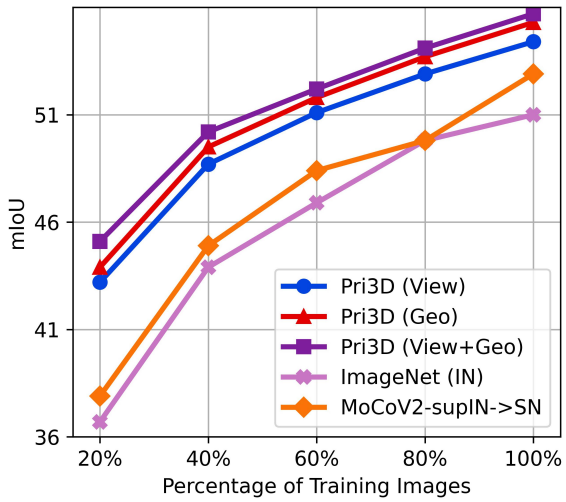

Figure 1: **Data Efficient Learning on ScanNet 2D Semantic Segmentation Task.** Our method achieves consistently better results with a ResNet18 backbone.

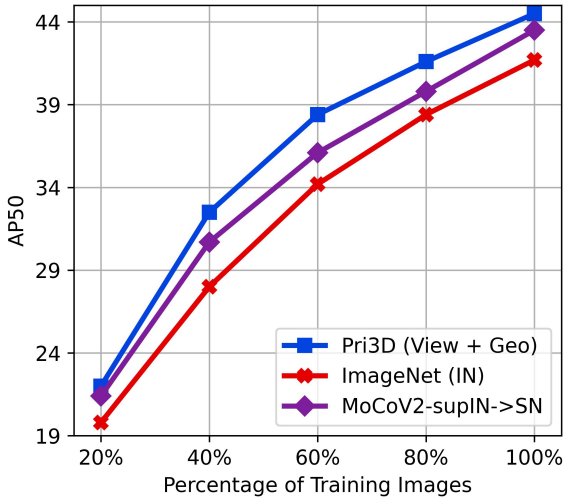

Figure 2: **Data Efficient Learning on ScanNet 2D Detection Task.** Similar to other tasks, our pre-training algorithm achieves consistently better results across various amounts of training data available. The backbone used is ResNet50.

To demonstrate that our pre-training algorithm generalizes well across different downstream tasks in limited data scenarios, we further plot data-efficient learning curves on the 2D object detection task on ScanNet in Figure 2, as well as data-efficient learning curves on the 2D instance segmentation task in Figure 3. We use Mask-RCNN [6] with a ResNet50 backbone for both tasks.

2. Additional Qualitative Visualizations

We additionally show more visualizations of 2D semantic segmentation on the ScanNet [3] and NYUv2 [11] datasets (see Figure 4). We can achieve notably improved segmentation results by using our pre-trained weights than ImageNet pre-trained weights.

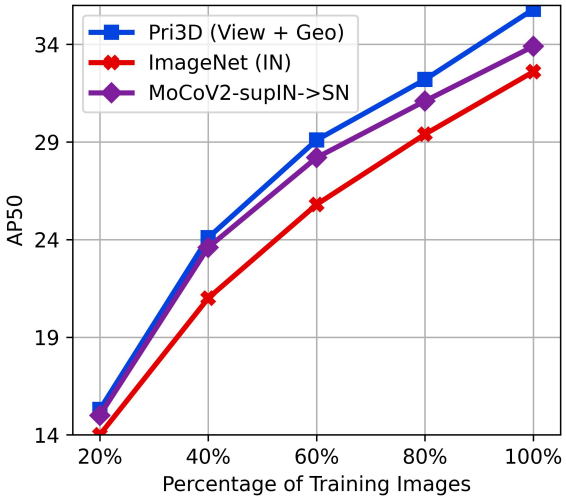

Figure 3: **Data Efficient Learning on ScanNet 2D Instance Segmentation Task.** Our approach provides improved performance across varying amounts of train images available. We use a ResNet50 backbone.

3. Generalization Across Datasets.

We conduct additional experiments on more diverse datasets for pre-training (ScanNet, MegaDepth [9], KITTI [5]) and downstream tasks including COCO [10] and outdoor data (Cityscapes [2] and KITTI) (see Table 1). We still observe a consistent improvement with our pre-training methods across different datasets and tasks. On COCO, the gap is not as significant as it in other scenarios, due to the drastic domain gap between COCO and our 3D dataset for pre-training.

Pretraining	Finetuning	Task	Metric
ImageNet-Pretrain	COCO	det (mAP)	59.5
Pri3D (ScanNet)	COCO	det (mAP)	**60.6**
ImageNet-Pretrain	COCO	ins (mAP)	56.6
Pri3D (ScanNet)	COCO	ins (mAP)	**57.5**
ImageNet-Pretrain	Cityscapes	sem (mIoU)	54.1
Pri3D (MegaDepth)	Cityscapes	sem (mIoU)	55.2
Pri3D (KITTI)	Cityscapes	sem (mIoU)	**56.3**
ImageNet-Pretrain	KITTI	sem (mIoU)	28.5
Pri3D (MegaDepth)	KITTI	sem (mIoU)	30.8
Pri3D (KITTI)	KITTI	sem (mIoU)	**33.2**

Table 1: Mask R-CNN is used for detection (det) and instance segmentation (ins); ResUNet50 for semantic segmentation (sem).

4. Unsupervised Pre-training Pipeline

We experiment with network weights trained with self-supervision on ImageNet for encoder initialization. This is also a two-stage pipeline but without using any semantic labels in either stage. Even though the use of a supervised ImageNet pre-trained initialization is a common practice, for completeness we also evaluate Pri3D in an unsupervised pipeline without using ImageNet labels; we demonstrate the experimental results of **Unsupervised Pri3D** in Table 2. Results suggest that Pri3D does not rely on any semantic supervision (*e.g.* ImageNet labels) to succeed, and still is able to achieve a substantial gain in this setup.

To clarify the baselines:

- **Unsupervised ImageNet Pre-training (MoCoV2-IN)**. We use MoCoV2 [1] ImageNet pre-trained weights. *No ScanNet data is involved.*

- **2-Stage MoCoV2 (MoCoV2-unsupIN→SN)**. We start with MoCoV2-IN as the encoder initialization, but add another stage to finetune MoCoV2 with randomly shuffled ScanNet images. In this case, *we use ScanNet data but no 3D priors are used.*

Method	mIoU
Scratch	39.1
ImageNet Pre-training (IN)	55.7
MoCoV2-IN	54.6 (-1.1)
MoCoV2-unsupIN→SN	55.4 (-0.3)
Unsupervised Pri3D (View)	60.5 (+4.8)
Unsupervised Pri3D (Geo)	60.8 (+5.1)
Unsupervised Pri3D (View + Geo)	**60.8** (+5.1)

Table 2: **2D Semantic Segmentation on ScanNet (Unsupervised Pre-training Pipeline).** We additionally show that for Pri3D encoder initialization (stage I), we can replace the ImageNet pre-trained weights with (self-supervised) MoCoV2 weights; the whole pipeline does not require semantic labels. Pri3D still shows a large improvement over supervised ImageNet pre-training and compare favorably with strong MoCo-style baselines. All experiments are with a ResNet50 backbone.

5. Limitations

While our approach demonstrates the promising effect of learning 3D priors for 2D representation learning, there are various limitations. For instance, joint 2D and 3D pre-training, in contrast to our current 3D-based constraints only, would likely provide the most informative signal for representation learning for downstream tasks. Additionally, our current 3D-based pre-training leverages indoor scene data from ScanNet, and we would expect further generalizability by augmentation with data from other environments, such as outdoor scene data (e.g., [4, 2]).

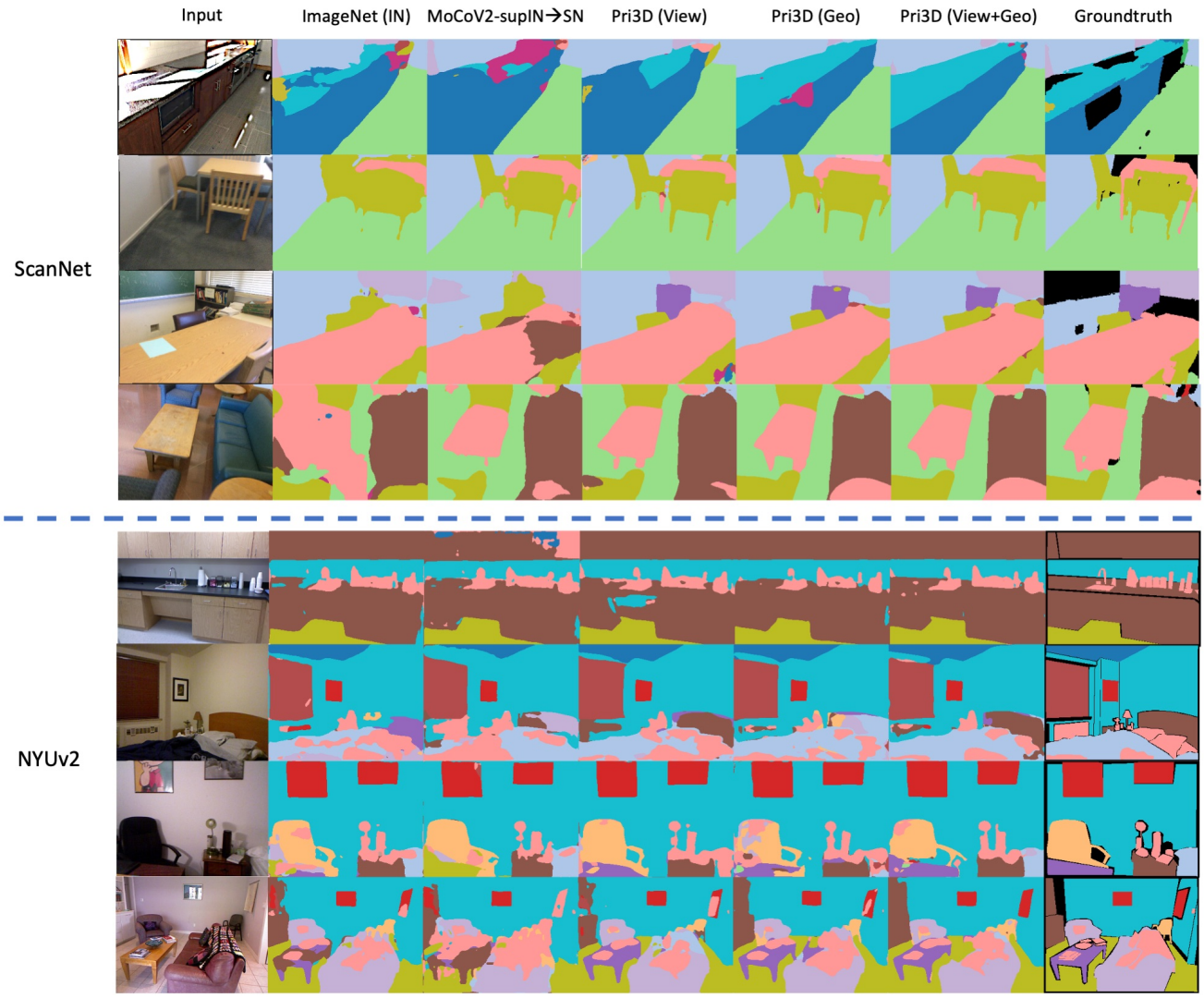

Figure 4: We show qualitative results on 2D semantic segmentation of ScanNet [3] and NYUv2 [11]. By encoding 3D priors, we obtain improved segmentation results, particularly for objects that are occluded or have more specular material properties.

References

[1] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017. 1, 3

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[8] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *CVPR*, 2021. 1

[9] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[10] 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 *ECCV*, 2014. 2

[11] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGB-D images. *ECCV*, 2012. 1, 3