Mask3D: Pre-training 2D Vision Transformers by Learning Masked 3D Priors
Supplementary Material

Ji Hou[1] Xiaoliang Dai[1] Zijian He[1] Angela Dai[2] Matthias Nießner[2]
[1]Meta Reality Labs [2]Technical University of Munich

1. More Qualitative Results

In this section, we show more visualizations on NYUv2 [6] and ScanNet [3] semantic segmentation results across different methods in Figure 1.

2. More Quantitative Results.

In this section, we show more quantitative results, including the full list of ablation studies regarding different depth and RGB ratios. Next, we show the results using MAE unsupervised pre-trained model against supervised pre-trained checkpoint for Stage-I pre-training. Furthermore, we show another out-of-domain transfer learning experiment on ADE20K, a more generally distributed dataset.

Full List of RGB and Depth Ratios. We show the expanded version of Table 8 below. We ablated different RGB and depth ratios, and found out that masking more RGB signal and bringing in sparse depth priors lead to higher mIoUs.

Out-of-domain Transfer on ADE20K. We observe a similar trend in the ADE20K [7] dataset compared to the ScanNet, NYUv2 and Cityscapes [2] (see the following Table 2). We search for the best training recipes: learning rate 0.0001 with AdamW optimizer, training iterations 256k and batch size 16 on 8 GPUs.

MAE-unsup-ViT vs. supIN-ViT. We list the baselines with ImageNet supervised pre-training in the following Table 3. In the main paper, we did not include supIN - ViT, since MAE-unsupIN - ViT shows a better performance than supIN - ViT (in the MAE [4] paper), and MAE-unsupIN - ViT weights are readily available from official MAE codebase whereas supIN - ViT is not. Note that our method also uses MAE-unsupIN - ViT as a default initialization so it is a fair comparison, and this makes our method a pure unsupervised approach. Similar trend is observed as well in Pri3D [5].

RGB Ratio	Depth Ratio	mIoU
20.0%	0.0%	65.2
20.0%	20.0%	**66.7**
20.0%	50.0%	66.4
20.0%	80.0%	65.5
20.0%	100.0%	65.3
50.0%	0.0%	66.0
50.0%	20.0%	65.9
50.0%	50.0%	64.7
50.0%	80.0%	65.4
50.0%	100.0%	65.7
80.0%	0.0%	64.4
80.0%	20.0%	64.8
80.0%	50.0%	64.8
80.0%	80.0%	64.9
80.0%	100.0%	65.0
100.0%	0.0%	64.6
100.0%	20.0%	64.8
100.0%	50.0%	64.5
100.0%	80.0%	64.2
100.0%	100.0%	64.5

Table 1. **Full List of RGB and Depth Ratios** Results on ScanNet 2D semantic segmentation. We mask out different ratios of RGB and depth patches, where the ratio indicates the percentage of kept patches.

Pre-training Method	Backbone	mIoU
MAE (MultiMAE reproduced) [1]	ViT	46.2
MAE (our reproduced)	ViT	47.2
MultiMAE [1]	ViT	46.2
Mask3D	ViT	47.7

Table 2. **Out-of-domain Transfer on ADK20k semantic segmentation**. Mask3D and MAE use the same training recipe for the downstream task, so it is a fair comparison. We can observe an improvement over MAE pre-trained checkpoint with masked depth priors for pre-training.

Limitations. Our work aims to learn 3D geometric and spatial structures to benefit downstream scene understand-

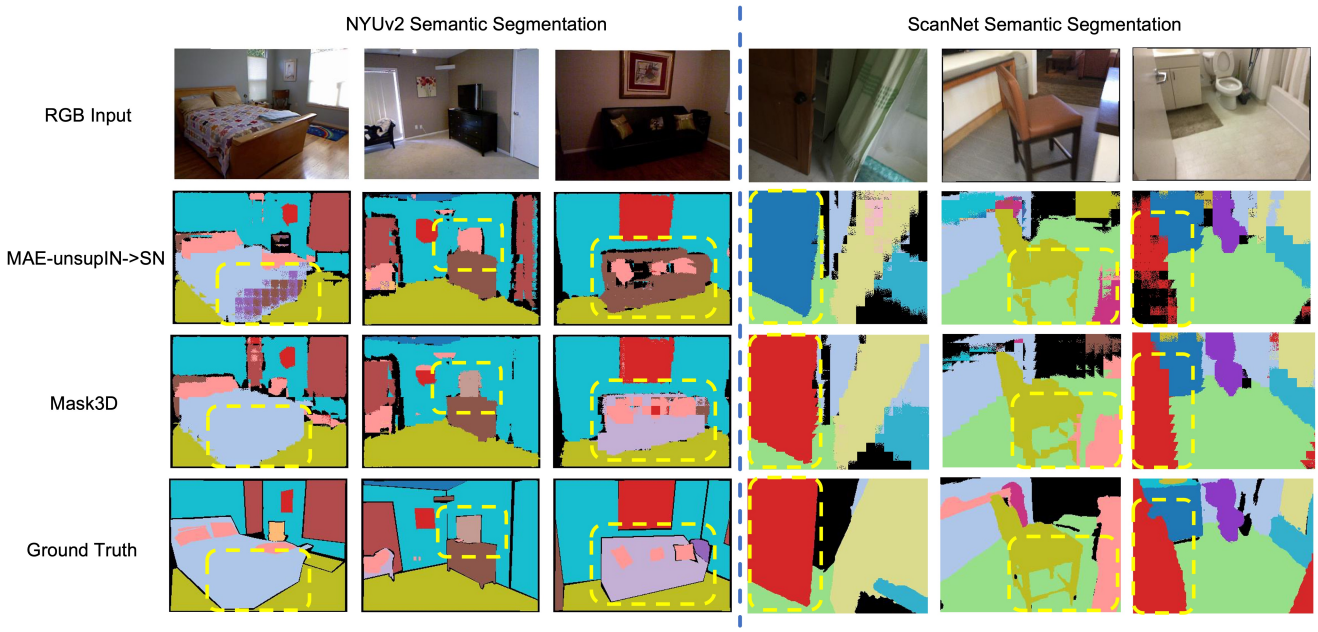

Figure 1. **Qualitative Results.** We show more visualizations on NYUv2 and ScanNet.

Method	mIoU
ViT Scratch	32.6
MAE-unsupIN - ViT	63.7
Mask3D - ViT	66.7

Table 3. **ViT baselines on ScanNet Semantic Segmentation.** A similar trend is observed in unsupervised setup.

ing tasks. While we show that learning to reconstruct the dense depth can effectively help embed learned geometric understanding, some geometric- and spatially-aware designs are not yet fully exploited, e.g., ViT-based multi-scale learning or exploring surface properties such as normals as a proxy loss, etc.

Training and Validation Curves. We demonstrate the training and validation curves of fine-tuning ScanNet Semantic Segmentation in Figure 2. A consistent improvement can be observed from the curve.

Pre-training Orders. We ablate the orders of pre-training in Table 4.

Pre-training Orders	mIoU
MAE + Mask3D	63.3
MAE $\rightarrow$ Mask3D	66.3
Mask3D $\rightarrow$ MAE	63.1

Table 4. **ScanNet 2D Semantic Segmentation.** "+" indicates training together and "$\rightarrow$" indicates the pre-training order.

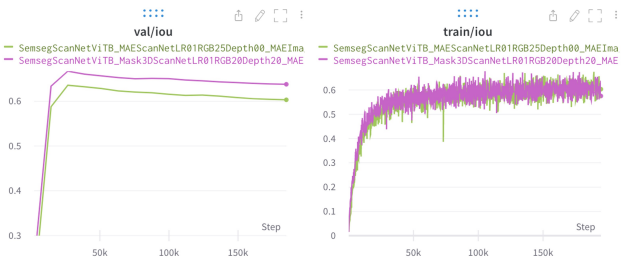

Figure 2. **Training and Validation Curves.** A consistent gap is observed on ScanNet Semantic Segmentation between Mask3D and MAE-unsupIN$\rightarrow$SN.

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 348–367. Springer, 2022. 1

[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. 1

[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

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 1

[5] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *ICCV*, 2021. 1

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

[7] 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. 1