2D-3D Interlaced Transformer for Point Cloud Segmentation with Scene-Level Supervision Supplementary Material

This document provides additional experiments. In the following, we first present some potential extensions of the proposed method. Then, we provide the model architecture and running time. Lastly, we present detailed quantitative results as well as qualitative examples, including less successful cases.

1. Extensions of the Proposed Method

1.1. Extension with Known Poses and Depths

In Section 4.3.2 of the submitted paper, we present how to extend our method when camera poses and depth maps are available. In the following, we elaborate on this extension by providing further details.

Inspired by the fact that positional information can enrich image features [5], we perform positional embedding for both 2D and 3D features before passing them to the transformer encoders. This way, both 2D and 3D share a common 3D world space, facilitating explicit position correlation between 2D images and 3D point clouds. We first generate the 3D coordinate map $\mathbf{x}_t \in \mathbb{R}^{H\times W\times 3}$ for each view $\mathbf{v}_t$. Given the depth map $\mathbf{d}_t$ and camera projection matrix $\mathbf{k}_t$, the 3D world coordinate $\mathbf{x}_t(u,v)$ at 2D position $[u,v]$ is computed by

$$[(\mathbf{x}_t(u,v))^\top, 1]^\top = \mathbf{d}_t(u,v)\cdot \mathbf{k}_t^{-1}[u,v,1]^\top. \quad (1)$$

Via Eq. 1, we obtain the 3D world coordinate map $\mathbf{x}_t$ for each view image $\mathbf{v}_t$. All T 3D coordinate maps $\{\mathbf{x}_t\}_{t=1}^T$ are fed into a coordinate embedding module f_{emb}, which is composed of two 1×1 convolution layers with ReLU activation, to get the positional embedding $z_{\text{2D}} \in \mathbb{R}^{T\times H\times W\times D}$, where D is the embedding dimension. The positional embedding is added to 2D features for 3D positional awareness. Since each point of a point cloud P lies in the 3D space, f_{emb} is directly applied to all points and gets $z_{\text{3D}} \in \mathbb{R}^{M\times D}$, where M is the number of points in P. Figure 1 depicts the extended method by revising Figure 2 in the submitted paper.

1.2. Extension to Joint 2D-3D Segmentation

As discussed in Section 5 of the submitted paper, the proposed method can be extended to joint 2D and 3D segmentation using weak supervision. Through flattening the image features F_{2D} instead of applying global average pooling, we obtain a set of multi-view patch tokens, which can be further considered as segmentation results [8]. However, we get inferior results, 0.275 and 0.129 in mIoU of 3D and 2D, respectively, on the ScanNet training set. It may be because too many views generate many patch tokens and create lots of noise. Specifically, the image size is 320×240 in ScanNet, creating 80 patch tokens for a view and a total of 1280 tokens for 16 views. The high number of 2D tokens hinders the optimization of self-attention and cross-attention, leading to unsatisfactory performance.

2. Model Architecture and Running time

As mentioned in Section 3.5, ResNet-50 [2] is adopted as the 2D feature extractor. MinkowskiNet [1] work as the 3D feature extractor for ScanNet and S3DIS. Specifically, we use MinkowskiUNet18A, and the voxel size is set to 5cm. The network was optimized on a machine with eight NVIDIA GTX 3090 GPUs. With 500 epochs, it took about two days to complete the optimization. For the semantic segmentation model, we use MinkowskiUNet18C with the voxel size set to 2cm. The network was optimized on a machine with eight NVIDIA GTX 1080Ti GPUs. The optimization required 150 epochs, which took around a day to complete.

As discussed in Sec 4.2.2, the overhead of the proposed interlaced decoder is acceptable. Table 1 shows the inference time and computational cost of different methods. WYPR [7] is not presented since their model code are not publicly available.

2.1. Generalize to Another Backbone

To assess the generalizability of our approach, we employ PointNet++ [6] as the 3D feature extractor, which also serves as a popular backbone for point cloud-based applications. Notably, our method yields competitive performance

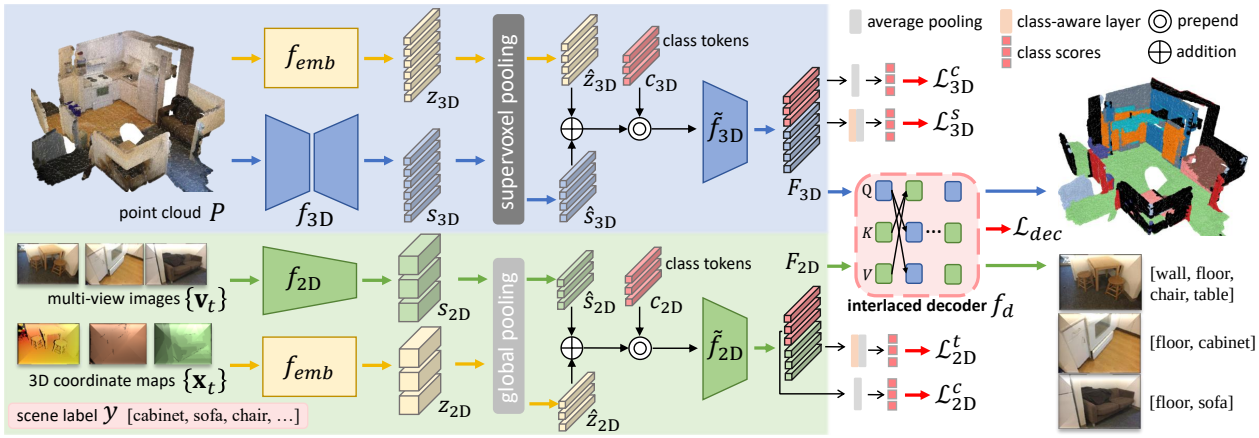

Figure 1: Network architecture of our MIT extension with camera poses and depths maps.

Methods	Time	FLOPs
MIL-Trans [9]	8.9 ms	181 G
MIT (3D-only)	9.4 ms	199 G
MIT (Ours)	27.3 ms	220 G

Table 1: The inference time and FLOPs of different methods.

results, achieving 35.1 and 29.7 mIoU on the ScanNet validation set and S3DIS test set, respectively. The results demonstrate that our MIT is general because it can work with different backbones.

2.2. Utilizing no pre-trained 2D model

We train our method with randomly initialized 2D ResNet-50 and observe only a minor performance drop (35.8% → 34.6% in mIoU on the ScanNet validation set). This result indicates that our method does not rely heavily on ImageNet pre-training and can work with pure 3D scene-level supervision.

3. Quantitative Results of Multi-view Images

We also report the performance of multi-label image classification in Table 2. This task aims to find all existing categories in a single view, by giving only the class appearance in the multi-view images for training. We first report the baseline result, which is conducted by averaging the estimated class scores across all views during training and obtaining the per-view classification result by passing a single view to the model. ResNet-50 [3] is adopted as the feature extractor. Our method enriches the views with self-attention in 2D and cross-attention in 3D, showing better classification results compared to competing methods that only consider 2D information.

Method	Sup.	ScanNet	S3DIS
Baseline	$\mathcal{F}$.	79.4	82.3
Baseline	$\mathcal{S}$.	52.4	55.1
Kim *et al.* [4]	$\mathcal{S}$.	54.9	58.2
MIT (Ours)	$\mathcal{S}$.	56.1	57.9

Table 2: Quantitative results (mAP) of several multi-label classification methods with diverse supervision settings on the ScanNet and S3DIS image datasets. "Sup." denotes the type of supervision. "$\mathcal{F}$." represents full annotation for each view. "$\mathcal{S}$." indicates that class tag annotation is shared by all views in the scene.

4. Class-wise Quantitative Performance

In Table 1 of the submitted paper, we report the average performance over the categories of ScanNet and S3DIS. Here we provide detailed class-wise results. Table 3, Table 4 and Table 5 show the class-wise performance of our method on the ScanNet validation set, ScanNet test set, and S3DIS datasets, respectively.

5. Qualitative Results

Figure 2 and Figure 3 show the segmentation results generated by our method with scene-level supervision, including both successful cases and less successful ones. It can be observed that the proposed method delineates precise segmentation contours without using any point-level supervision. However, those categories with very similar shapes and colors lead to wrong segmentation results, such as the other furniture in the example of the last column of the third row in Figure 2 and the clutter in the example of the last column of the first row in Figure 3. Also, some points of wall examples may be misclassified as doors or windows since they share very similar shapes.

Method	wall	floor	cabinet	bed	chair	sofa	table	door	window	B.S.	picture	cnt	desk	curtain	fridge	S.C.	toilet	sink	bathtub	other	mIOU
MIL-trans [9]	52.1	50.6	8.3	46.3	27.9	39.7	20.9	15.8	26.8	40.2	8.1	21.1	22.0	45.9	4.5	16.6	15.2	32.4	21.2	8.0	26.2
WYPR [7]	52.0	77.1	6.6	54.3	35.2	40.9	29.6	9.3	28.7	33.3	4.8	26.6	27.9	69.4	8.1	27.9	24.1	25.4	32.3	8.7	31.1
MIT (Ours)	57.3	89.7	24.1	54.9	31.5	62.8	42.5	19.8	27.4	45.1	1.1	31.4	41.7	41.4	17.6	25.0	34.5	8.3	44.4	15.6	35.8

Table 3: Quantitative results (mIoU) of several point-cloud segmentation methods with scene-level supervision setting on the ScanNet validation set. "B.S." denotes bookshelf; "S.C." stands for shower curtain and "cnt" denotes counter.

Method	wall	floor	cabinet	bed	chair	sofa	table	door	window	B.S.	picture	cnt	desk	curtain	fridge	S.C.	toilet	sink	bathtub	other	mIOU
MIT (Ours)	42.2	82.1	16.3	55.8	30.6	57.6	35.9	19.3	27.0	39.0	1.4	25.3	27.7	31.3	21.3	17.8	47.8	7.9	29.8	18.8	31.7

Table 4: Quantitative results (mIoU) of our method with scene-level supervision setting on the test set from official ScanNet benchmark server. "B.S." denotes bookshelf; "S.C." stands for shower curtain and "cnt" denotes counter.

Supervision	ceil	floor	wall	beam	column	window	door	chair	table	bookcase	sofa	board	clutter	mIOU
MIL-Trans [9]	24.9	4.7	40.0	0.0	1.3	2.2	1.8	5.6	16.8	33.0	32.1	0.1	5.8	12.9
MIT (Ours)	80.8	81.0	81.8	0.0	0.9	0.2	27.6	26.7	19.5	15.5	16.8	0.0	9.9	27.7

Table 5: Quantitative results (mIoU) of the proposed method with diverse supervision settings on the S3DIS Area 5 dataset.

References

[1] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[4] Youngwook Kim, Jae Myung Kim, Zeynep Akata, and Jungwoo Lee. Large loss matters in weakly supervised multi-label classification. In *CVPR*, 2022. 2

[5] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3D object detection. In *ECCV*, 2022. 1

[6] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, 2017. 1

[7] Zhongzheng Ren, Ishan Misra, Alexander G Schwing, and Rohit Girdhar. 3D spatial recognition without spatially labeled 3D. In *CVPR*, 2021. 1, 3

[8] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *CVPR*, 2022. 1

[9] Cheng-Kun Yang, Ji-Jia Wu, Kai-Syun Chen, Yung-Yu Chuang, and Yen-Yu Lin. An mil-derived transformer for weakly supervised point cloud segmentation. In *ICCV*, 2022. 2, 3

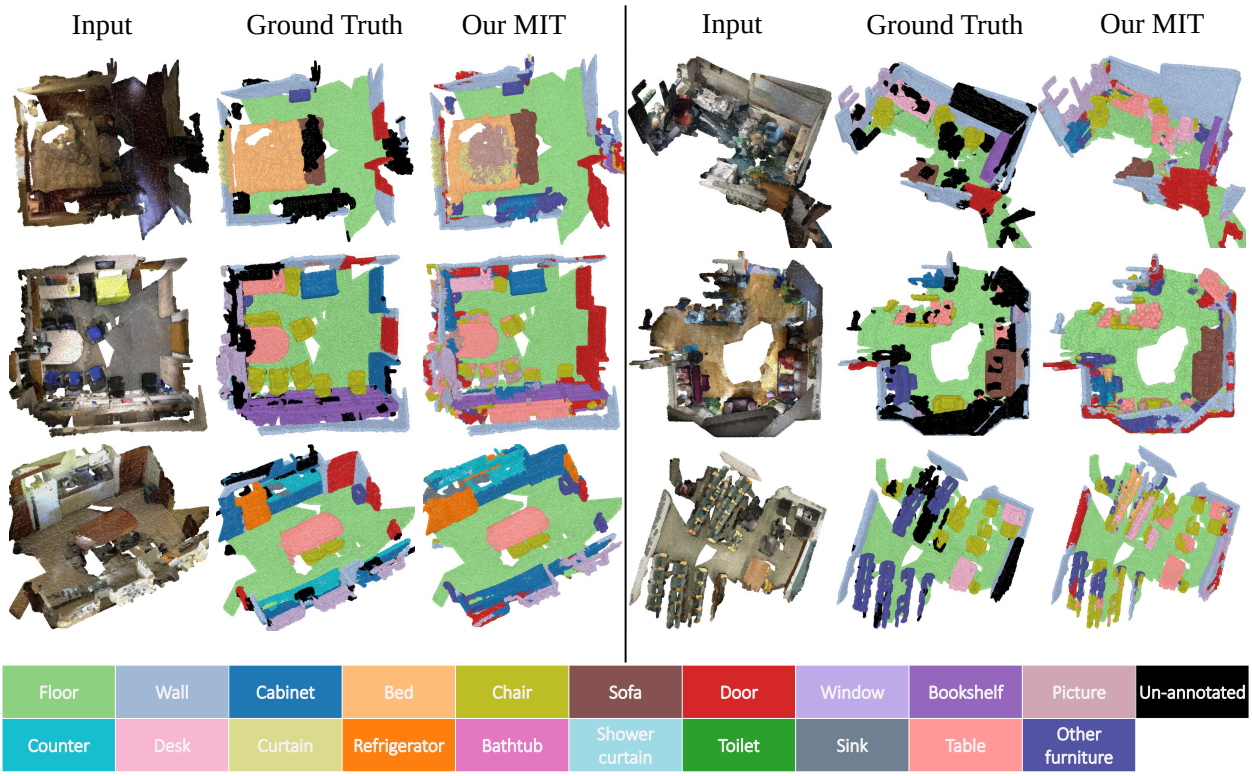

Figure 2: Qualitative results on the ScanNet dataset with scene-level supervision. Each category is associated with the same row. For each example, we show the input cloud, the ground-truth label, and our segmentation result. The last example of each row (on the right of the gray line) shows a less successful case.

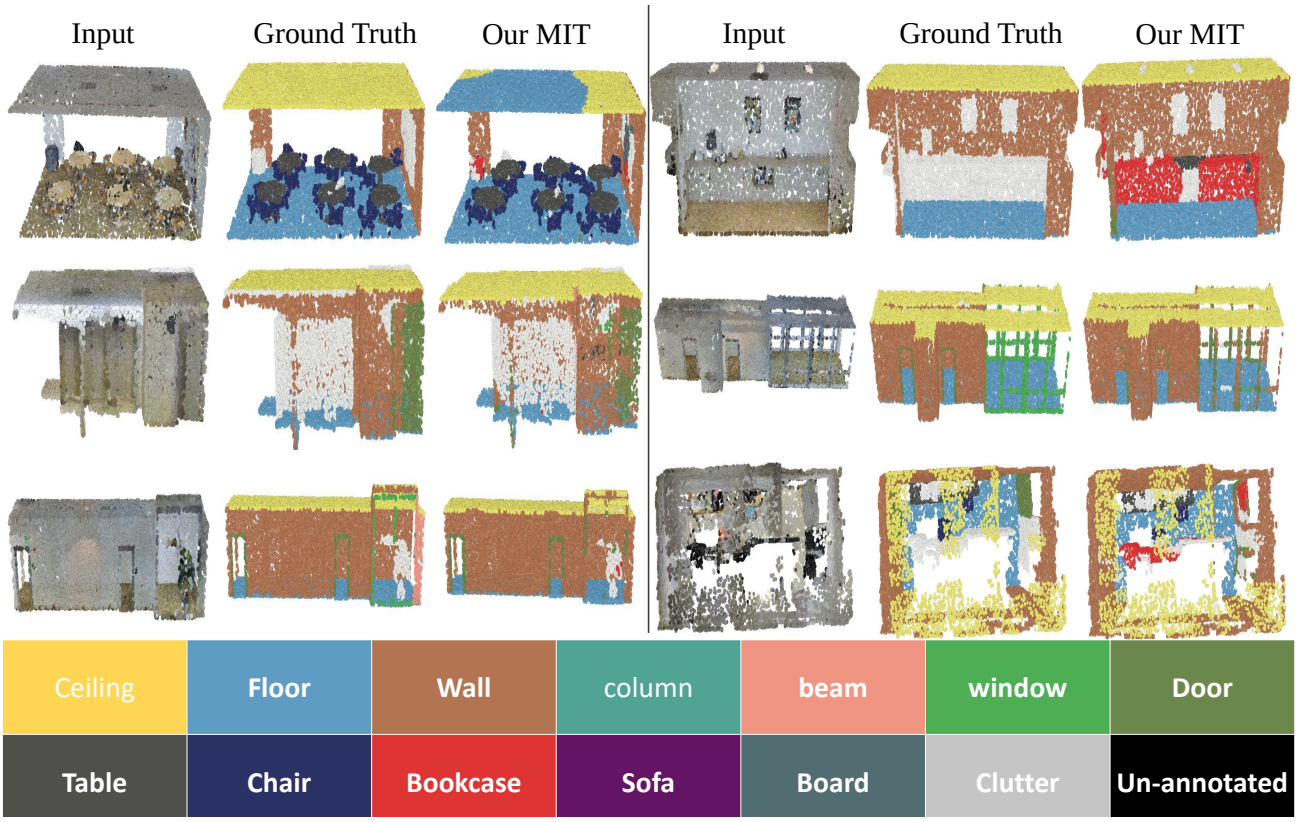

Figure 3: Qualitative results on the S3DIS dataset with scene-level supervision. Each category is associated with the same row. For each example, we show the input cloud, the ground-truth label, and our segmentation result. The last example of each row (on the right of the gray line) shows a less successful case.