Relation3D: Enhancing Relation Modeling for Point Cloud Instance Segmentation

Supplementary Material

1. Overview

In this supplementary material, we begin by presenting a more detailed comparison of quantitative metrics on ScanNetV2 [1] validation set and test set (Section 2). We then provide additional discussion about position embedding (Section 3). To further validate the effectiveness of the proposed method, we provide more visualizations (Section 5).

2. Detailed results on ScanNetV2 validation and hidden test set

The detailed results for each category on ScanNetV2 validation set are reported in Table 1. As the table illustrates, our method achieves the best performance in 14 out of 18 categories. The two of them work together to achieve 14 out of 18 categories. The superior performance demonstrates the effectiveness of our method. The detailed results for each category on ScanNetV2 hidden test set are reported in Table 2, 3 and 4. As the tables illustrate, our method achieves the best performance in 13 out of 18 categories in Table 2, 10 out of 18 categories in Table 3 and 12 out of 18 categories in Table 4. The superior performance demonstrates the effectiveness of our method.

3. Discussion about position embedding

In DETR-based methods, query typically consist of two embeddings: a content embedding and a position embedding. In the transformer decoder, the position embedding is added to the content embedding and then input into the self-attention/cross-attention mechanisms for interaction, ultimately generating a new content embedding. It is important to note that in these methods, the position represented by the position embedding does not accurately correspond to the actual location of the mask predicted by the query. In the following sections, we will analyze several representative methods in detail.

SPFormer: In SPFormer, both the content embedding and position embedding are learnable. Consequently, the position embedding does not carry any explicit spatial meaning.

Mask3D: In Mask3D, the position embedding is derived using FPS (Furthest Point Sampling). First, N points are sampled using FPS, and each sampled point is encoded using Fourier or sin-cos encoding to generate the corresponding position embedding. The content embedding, however, is initialized to all zeros. During the subsequent decoder process, Mask3D employs self-attention and cross-attention mechanisms to update the features of the content embedding. However, through experiments, we observe that the positions of the sampling points in Mask3D do not align with the actual positions of the corresponding predicted instances (there are large differences between them). This indicates that the positional relationships introduced by the position embedding are inaccurate.

Maft: Maft also employs a learnable position embedding, but unlike SPFormer, the position embedding $P \in [0, 1]^{N\times 3}$. The position embedding is resized based on the scale of the input scene as follows:

$$\hat{P} = P \cdot (p_{\text{max}} - p_{\text{min}}) + p_{\text{min}},$$

where $p_{\text{max}}, p_{\text{min}} \in \mathbb{R}^3$ denote the maximum and minimum coordinates of the input scene, respectively. As a result, the position embedding in Maft acquires actual coordinate meanings. To correlate the position embedding $\hat{P}$ more closely with the positions of the corresponding predicted instances, Maft introduces a C_{center} term in the Hungarian matching cost matrix, representing the distance between $\hat{P}$ and the ground truth instance center. Furthermore, Maft updates $\hat{P}$ layer by layer in the decoder, allowing the matched $\hat{P}$ to progressively approach the ground truth instance center. Despite these design enhancements, where $\hat{P}$ becomes closer to the actual positions of the corresponding predicted instances, some discrepancies remain. These inaccuracies impact the precision of the positional relationships.

4. Parameter and Runtime Analysis.

Table 5 presents model parameters and runtime per scan for various methods evaluated on ScanNetV2 validation set. For a fair comparison, all runtimes are measured on the same RTX 4090 GPU. Compared to Maft, our method achieves better performance with an additional 2.0M parameters. Although our method is slightly slower due to additional modules, the small-scale design allows our approach to outperform most methods in both speed and parameter efficiency.

5. More Visualization

Qualitative comparison (Figure 1): To vividly illustrate the differences between our method and baseline, we visualize qualitative results in Figure 1. From the regions highlighted in the last row, we observe that the baseline method tends to confuse chairs with surrounding objects and exhibits incomplete segmentation of the chair. In contrast, our method, by focusing on scene feature modeling, enhances the consistency of superpoint features within instances and increases the differences between features of different instances. This leads to more accurate and coherent segmentation results.

Method	mAP	bathtub	bed	bookshe.	cabinet	chair	counter	curtain	desk	door	other	picture	frige	s. curtain	sink	sofa	table	toilet	window
PointGroup [2]	34.8	59.7	37.6	26.7	25.3	71.2	6.9	26.6	14.0	22.9	33.9	20.8	24.6	41.6	29.8	43.4	38.5	75.8	27.5
SSTNet [3]	49.4	77.7	56.6	25.8	40.6	81.8	22.5	38.4	28.1	42.9	52.0	40.3	43.8	48.9	54.9	52.6	55.7	92.9	34.3
SoftGroup [4]	45.8	66.6	48.4	32.4	37.7	72.3	14.3	37.6	27.6	35.2	42.0	34.2	56.2	56.9	39.6	47.6	54.1	88.5	33.0
DKNet [5]	50.8	73.7	53.7	36.2	42.6	80.7	22.7	35.7	35.1	42.7	46.7	51.9	39.9	57.2	52.7	52.4	54.2	91.3	37.2
Mask3D [6]	55.2	78.3	54.3	43.5	47.1	82.9	35.9	48.7	37.0	54.3	59.7	53.3	47.7	47.4	55.6	48.7	63.8	94.6	39.9
ISBNet [7]	54.5	76.3	58.0	39.3	47.7	83.1	28.8	41.8	35.9	49.9	53.7	48.6	51.6	66.2	56.8	50.7	60.3	90.7	41.1
SPFormer [8]	56.3	83.7	53.6	31.9	45.0	80.7	38.4	49.7	41.8	52.7	55.6	55.0	57.5	56.4	**59.7**	51.1	62.8	**95.5**	41.1
QueryFormer [9]	56.5	81.3	57.7	**45.0**	47.2	82.0	37.2	43.2	43.3	54.5	60.5	52.6	54.1	62.7	52.4	49.9	60.5	94.7	37.4
Maft [10]	58.4	80.1	58.1	41.8	48.3	82.2	34.4	**55.1**	44.3	55.0	57.9	61.6	56.4	63.7	54.4	53.0	66.3	95.3	42.9
Ours	**62.5**	**84.8**	**63.7**	42.2	**54.0**	**83.9**	**49.1**	53.8	**46.7**	**60.4**	**65.0**	**62.8**	**61.3**	**69.2**	57.6	**61.2**	**68.9**	94.3	**46.7**

Table 1. **Full quantitative results of mAP on ScanNetV2 validation set.** For reference purposes, we show the results of fully supervised methods in gray. Best performance of box supervised methods is in boldface.

Method	mAP	bathtub	bed	bookshe.	cabinet	chair	counter	curtain	desk	door	other	picture	frige	s. curtain	sink	sofa	table	toilet	window
3D-BoNet [11]	25.3	51.9	32.4	25.1	13.7	34.5	3.1	41.9	6.9	16.2	13.1	5.2	20.2	33.8	14.7	30.1	30.3	65.1	17.8
MTML [12]	28.2	57.7	38.0	18.2	10.7	43.0	0.1	42.2	5.7	17.9	16.2	7.0	22.9	51.1	16.1	49.1	31.3	65.0	16.2
3D-MPA [13]	35.5	45.7	48.4	29.9	27.7	59.1	4.7	33.2	21.2	21.7	27.8	19.3	41.3	41.0	19.5	57.4	35.2	84.9	21.3
DyCo3D [14]	39.5	64.2	51.8	44.7	25.9	66.6	5.0	25.1	16.6	23.1	36.2	23.2	33.1	53.5	22.9	58.7	43.8	85.0	31.7
PE [15]	39.6	66.7	46.7	44.6	24.3	62.4	2.2	57.7	10.6	21.9	34.0	23.9	48.7	47.5	22.5	54.1	35.0	81.8	27.3
PointGroup [2]	40.7	63.9	49.6	41.5	24.3	64.5	2.1	57.0	11.4	21.1	35.9	21.7	42.8	66.6	25.6	56.2	34.1	86.0	29.1
MaskGroup [16]	43.4	77.8	51.6	47.1	33.0	65.8	2.9	52.6	24.9	25.6	40.0	30.9	38.4	29.6	36.8	57.5	42.5	87.7	36.2
OccuSeg [17]	48.6	80.2	53.6	42.8	36.9	70.2	20.5	33.1	30.1	37.9	47.4	32.7	43.7	**86.2**	48.5	60.1	39.4	84.6	27.3
HAIS [18]	45.7	70.4	56.1	45.7	36.4	67.3	4.6	54.7	19.4	30.8	42.6	28.8	45.4	71.1	26.2	56.3	43.4	88.9	34.4
SSTNet [3]	50.6	73.8	54.9	49.7	31.6	69.3	17.8	37.7	19.8	33.0	46.3	57.6	51.5	85.7	**49.4**	63.7	45.7	94.3	29.0
SoftGroup [4]	50.4	66.7	57.9	37.2	38.1	69.4	7.2	67.7	30.3	38.7	53.1	31.9	58.2	75.4	31.8	64.3	49.2	90.7	38.8
DKNet [5]	53.2	81.5	62.4	51.7	37.7	74.9	10.7	50.9	30.4	43.7	47.5	58.1	53.9	77.5	33.9	64.0	50.6	90.1	38.5
Mask3D [6]	56.6	92.6	59.7	40.8	42.0	73.7	23.9	59.8	38.6	45.8	54.9	56.8	71.6	60.1	48.0	64.6	57.5	92.2	36.4
QueryFormer [9]	58.6	92.6	70.2	39.3	**50.5**	73.7	27.7	58.3	37.5	47.9	53.5	56.8	61.5	72.0	48.1	74.5	59.2	95.8	36.1
Maft [10]	59.6	88.9	**72.1**	44.8	46.0	76.8	25.1	55.8	40.8	50.4	53.9	61.6	61.8	85.8	48.2	68.4	55.1	93.1	**45.0**
Ours	**62.2**	**92.6**	71.0	**54.1**	50.2	**77.2**	**31.4**	**59.8**	**42.5**	**50.4**	**56.5**	**65.0**	**71.6**	80.9	47.6	**74.7**	**61.8**	**96.3**	36.4

Table 2. **Full quantitative results of mAP on the ScanNetV2 test set. Best performance is in boldface.**

T-SNE Visualization (Figure 2): We present the T-SNE visualization of superpoint-level feature distributions on the ScanNetV2 validation set. From this visualization, it is evident that our method achieves better inter-object differentiation while maintaining intra-object feature consistency.

Attention Maps and Weight Distributions (Figures 3 and 4): We compare traditional self-attention with relation-aware self-attention in terms of attention maps and weight distributions. Relation-aware self-attention exhibits a more focused attention mechanism, effectively modeling positional and geometric relationships. Unlike traditional self-attention, which distributes attention more diffusely without a clear focal query, relation-aware self-attention emphasizes relevant queries, leading to a more precise and meaningful feature representation.

Adaptive Superpoint Aggregation Module (Figure 5): Figure 5 visualizes the weights from the Adaptive Superpoint Aggregation Module (ASAM). The visualization highlights that ASAM assigns higher weights to object edges and corners—regions typically distinctive for individual instances. This targeted emphasis enhances the model's ability to accurately capture key structural features, contributing to better

Method	AP@50	bathtub	bed	bookshe.	cabinet	chair	counter	curtain	desk	door	other	picture	frige	s. curtain	sink	sofa	table	toilet	window
3D-BoNet [11]	48.8	100.0	67.2	59.0	30.1	48.4	9.8	62.0	30.6	34.1	25.9	12.5	43.4	79.6	40.2	49.9	51.3	90.9	43.9
MTML [12]	54.9	100.0	80.7	58.8	32.7	64.7	4	81.5	18.0	41.8	36.4	18.2	44.5	100.0	44.2	68.8	57.1	100.0	39.6
3D-MPA [13]	61.1	100.0	83.3	76.5	52.6	75.6	13.6	58.8	47.0	43.8	43.2	35.8	65.0	85.7	42.9	76.5	55.7	100.0	43.0
DyCo3D [14]	64.1	100.0	84.1	89.3	53.1	80.2	11.5	58.8	44.8	43.8	53.7	43.0	55.0	85.7	53.4	76.4	65.7	98.7	56.8
PE [15]	64.5	100.0	77.3	79.8	53.8	78.6	8.8	79.9	35.0	43.5	54.7	54.5	64.6	93.3	56.2	76.1	55.6	99.7	50.1
PointGroup [2]	63.6	100.0	76.5	62.4	50.5	79.7	11.6	69.6	38.4	44.1	55.9	47.6	59.6	100.0	66.6	75.6	55.6	99.7	51.3
MaskGroup [16]	66.4	100.0	82.2	76.4	61.6	81.5	13.9	69.4	59.7	45.9	56.6	59.9	60.0	51.6	71.5	81.9	63.5	100.0	60.3
OccuSeg [17]	67.2	100.0	75.8	68.2	57.6	84.2	47.7	50.4	52.4	56.7	58.5	45.1	55.7	100.0	75.1	79.7	56.3	100.0	46.7
HAIS [18]	69.9	100.0	84.9	82.0	67.5	80.8	27.9	75.7	46.5	51.7	59.6	55.9	60.0	100.0	65.4	76.7	67.6	99.4	56.0
SSTNet [3]	69.8	100.0	69.7	88.8	55.6	80.3	38.7	62.6	41.7	55.6	58.5	70.2	60.0	100.0	82.4	72.0	69.2	100.0	50.9
SoftGroup [4]	76.1	100.0	80.8	84.5	71.6	86.2	24.3	**82.4**	65.5	62.0	73.4	69.9	79.1	98.1	71.6	84.4	76.9	100.0	59.4
DKNet [5]	71.8	100.0	81.4	78.2	61.9	87.2	22.4	75.1	56.9	67.7	58.5	72.4	63.3	98.1	51.5	81.9	73.6	100.0	61.7
Mask3D [6]	78.0	100.0	78.6	71.6	69.6	88.5	50.0	71.4	**81.0**	67.2	71.5	67.9	80.9	100.0	**83.1**	83.3	78.7	100.0	60.2
QueryFormer [9]	78.4	100.0	93.3	60.1	**75.4**	88.5	56.4	67.7	66.6	66.4	71.6	67.9	**82.0**	100.0	83.0	89.7	80.4	100.0	62.2
Maft [10]	78.6	100.0	89.4	80.7	69.4	89.3	48.6	67.4	74.0	**78.6**	70.4	**72.7**	73.9	100.0	70.7	84.9	75.6	100.0	**68.5**
Ours	**81.6**	**100.0**	**97.1**	**90.8**	74.3	**92.3**	**57.3**	71.4	69.5	73.4	**74.7**	72.5	80.9	**100.0**	81.4	**89.9**	**82.0**	**100.0**	61.0

Table 3. **Full quantitative results of AP@50 on the ScanNetV2 test set. Best performance is in boldface.**

Method	AP@25	bathtub	bed	bookshe.	cabinet	chair	counter	curtain	desk	door	other	picture	frige	s. curtain	sink	sofa	table	toilet	window
3D-BoNet [11]	68.7	100.0	88.7	83.6	58.7	64.3	55.0	62.0	72.4	52.2	50.1	24.3	51.2	100.0	75.1	80.7	66.1	90.9	61.2
MTML [12]	73.1	100.0	99.2	77.9	60.9	74.6	30.8	86.7	60.1	60.7	53.9	51.9	55.0	100.0	82.4	86.9	72.9	100.0	61.6
3D-MPA [13]	73.7	100.0	93.3	78.5	79.4	83.1	27.9	58.8	69.5	61.6	55.9	55.6	65.0	100.0	80.9	87.5	69.6	100.0	60.8
DyCo3D [14]	76.1	100.0	93.5	89.3	75.2	86.3	60.0	58.8	74.2	64.1	63.3	54.6	55.0	85.7	78.9	85.3	76.2	98.7	69.9
PE [15]	77.6	100.0	90.0	86.0	72.8	86.9	40.0	85.7	77.4	56.8	70.1	60.2	64.6	93.3	84.3	89.0	69.1	99.7	70.9
PointGroup [2]	77.8	100.0	90.0	79.8	71.5	86.3	49.3	70.6	89.5	56.9	70.1	57.6	63.9	100.0	88.0	85.1	71.9	99.7	70.9
MaskGroup [16]	79.2	100.0	96.8	81.2	76.6	86.4	46.0	81.5	88.8	59.8	65.1	63.9	60.0	91.8	94.1	89.6	72.1	100.0	72.3
OccuSeg [17]	74.2	100.0	92.3	78.5	74.5	86.7	55.7	57.8	72.9	67.0	64.4	48.8	57.7	100.0	79.4	83.0	62.0	100.0	55.0
HAIS [18]	80.3	100.0	**99.4**	82.0	75.9	85.5	55.4	88.2	82.7	61.5	67.6	63.8	64.6	100.0	91.2	79.7	76.7	99.4	72.6
SSTNet [3]	78.9	100.0	84.0	88.8	71.7	83.5	71.7	68.4	62.7	72.4	65.2	72.7	60.0	100.0	91.2	82.2	75.7	100.0	69.1
SoftGroup [4]	86.5	100.0	96.9	86.0	86.0	91.3	55.8	**89.9**	91.1	76.0	**82.8**	73.6	80.2	98.0	91.9	87.5	87.7	100.0	82.0
DKNet [5]	81.5	100.0	93.0	84.4	76.5	91.5	53.4	80.5	80.5	80.7	65.4	76.3	65.0	100.0	79.4	88.1	76.6	100.0	75.8
Mask3D [6]	87.0	100.0	98.5	78.2	81.8	93.8	76.0	74.9	92.3	87.7	76.0	78.5	82.0	100.0	91.2	86.4	87.8	98.3	82.5
QueryFormer [9]	87.3	100.0	97.8	80.9	87.6	93.7	70.2	74.9	88.4	87.5	75.5	**78.5**	**83.5**	100.0	91.2	91.6	86.9	100.0	82.5
Maft [10]	86.0	100.0	99.0	81.0	82.9	94.9	80.9	68.8	83.6	90.4	75.1	79.6	74.1	100.0	86.4	84.8	83.7	100.0	**82.8**
Ours	**90.1**	**100.0**	97.8	**92.8**	**87.9**	**96.2**	**88.2**	74.9	**94.7**	**91.2**	80.2	75.3	82.0	**100.0**	**98.4**	**91.9**	**89.4**	**100.0**	81.5

Table 4. **Full quantitative results of AP@25 on the ScanNetV2 test set. Best performance is in boldface.**

performance in segmentation task.

6. More ablution study

We conduct an ablation study to analyze the effect of different values of r on model performance. By varying r, we evaluate the balance between computational efficiency and the quality of feature refinement. The results of this study are detailed in Table 6. Through this experiment, we observe that when $r = 3$, the performance is comparable to that of $r = 1$. However, considering that $r = 3$ (the interval of layers after which the refinement of F_{super} is performed is 3) reduces the frequency of feature refinement, the compu-

Method	Parameter(M)	Runtime(ms)
HAIS [18]	30.9	525
SSTNet [3]	/	663
SPFormer [8]	17.6	390
Mask3D [6]	39.6	525
SoftGroup [4]	30.9	535
Maft [10]	20.1	375
QueryFormer [9]	42.3	443
Spherical Mask [19]	30.8	432
Ours	22.1	394

Table 5. **Parameter and runtime analysis of different methods on ScanNetV2 validation set.** The runtime is measured on the same device.

r	ScanNetV2 validation			ScanNet200 validation		
	mAP	AP@50	AP@25	mAP	AP@50	AP@25
1	62.4	79.9	**87.1**	31.4	**41.3**	45.5
3	**62.5**	**80.2**	87.0	**31.6**	41.2	**45.6**
6	62.3	79.5	86.4	30.6	40.6	44.9

Table 6. **Ablation Study on r.** In our method, r represents the interval of layers after which the refinement of F_{super} is performed.

tational cost is consistently reduced. Thus, we set $r = 3$. Moreover, we find that the hyperparameter r demonstrates strong robustness, consistently achieving a good balance between accuracy and efficiency across different datasets.

7. Limitation and future work

Existing indoor 3D instance segmentation methods primarily focus on static objects and are typically performed offline (i.e., by first reconstructing point clouds from multiple RGB-D frames and then performing segmentation), which is not suitable for embodied environments. Therefore, future work needs to focus more on online instance segmentation in dynamic environments, where reconstruction and segmentation are not decoupled but rather performed simultaneously within a unified framework. A recent work, EmbodiedSAM [20], provides valuable insights that are worth exploring.

References

[1] 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.

[2] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020.

[3] Zhihao Liang, Zhihao Li, Songcen Xu, Mingkui Tan, and Kui Jia. Instance segmentation in 3d scenes using semantic superpoint tree networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2783–2792, 2021.

[4] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022.

[5] Yizheng Wu, Min Shi, Shuaiyuan Du, Hao Lu, Zhiguo Cao, and Weicai Zhong. 3d instances as 1d kernels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 235–252. Springer, 2022.

[6] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*, 2022.

[7] Tuan Duc Ngo, Binh-Son Hua, and Khoi Nguyen. Isbnet: a 3d point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13550–13559, 2023.

[8] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2393–2401, 2023.

[9] Jiahao Lu, Jiacheng Deng, Chuxin Wang, Jianfeng He, and Tianzhu Zhang. Query refinement transformer for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18516–18526, 2023.

[10] Xin Lai, Yuhui Yuan, Ruihang Chu, Yukang Chen, Han Hu, and Jiaya Jia. Mask-attention-free transformer for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3693–3703, 2023.

[11] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018.

[12] Jean Lahoud, Bernard Ghanem, Marc Pollefeys, and Martin R Oswald. 3d instance segmentation via multi-task metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9256–9266, 2019.

[13] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9031–9040, 2020.

[14] Tong He, Chunhua Shen, and Anton Van Den Hengel. Dyco3d: Robust instance segmentation of 3d point clouds through dynamic convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 354–363, 2021.

[15] Biao Zhang and Peter Wonka. Point cloud instance segmentation using probabilistic embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8883–8892, 2021.

[16] Min Zhong, Xinghao Chen, Xiaokang Chen, Gang Zeng, and Yunhe Wang. Maskgroup: Hierarchical point grouping

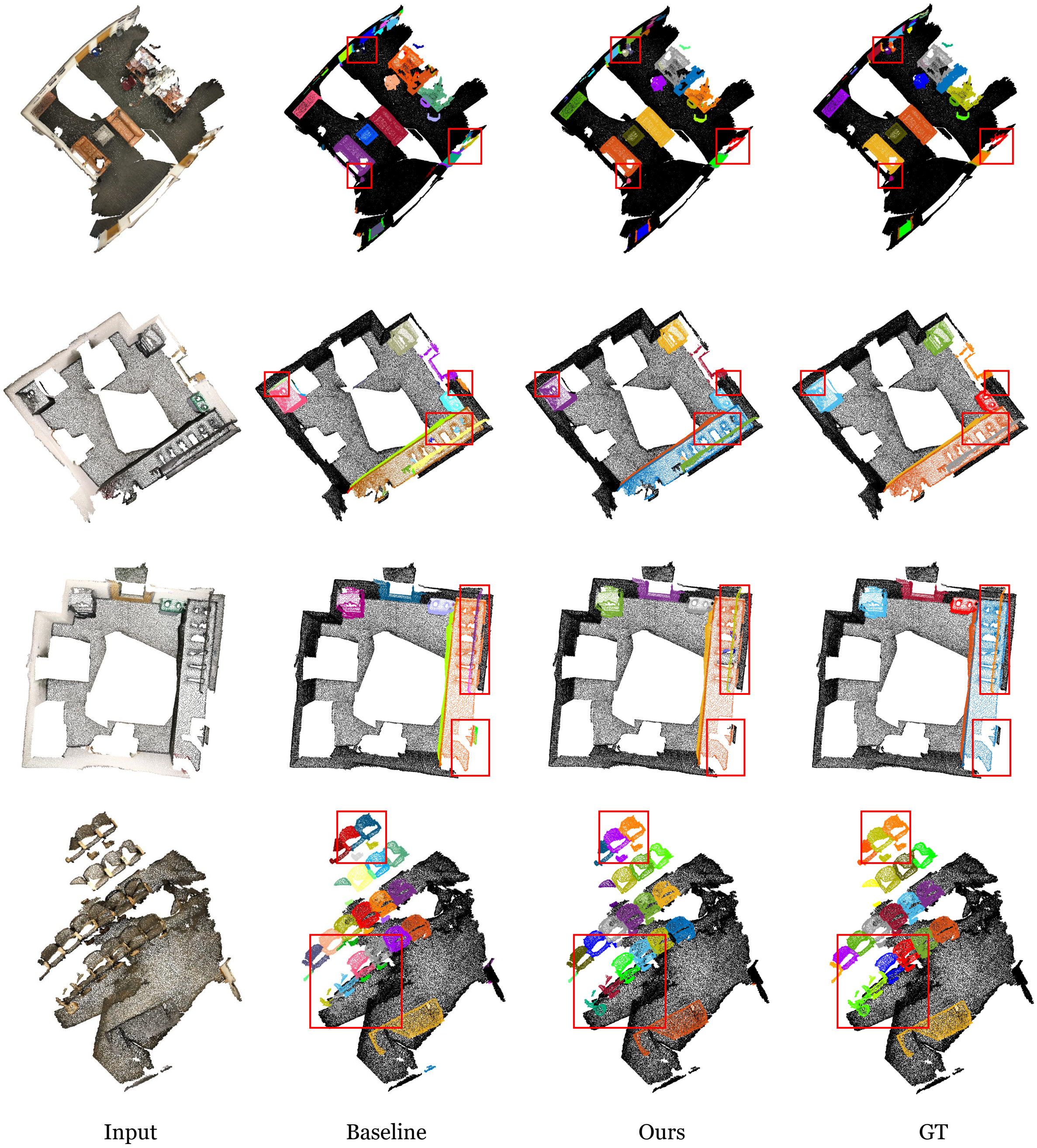

Figure 1. **More visualization of instance segmentation results on ScanNetV2 validation set.** The red boxes highlight the key regions

and masking for 3d instance segmentation. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

[17] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2940–2949, 2020.

[18] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15467–15476, 2021.

[19] Sangyun Shin, Kaichen Zhou, Madhu Vankadari, Andrew Markham, and Niki Trigoni. Spherical mask: Coarse-to-fine 3d point cloud instance segmentation with spherical representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4060–4069,

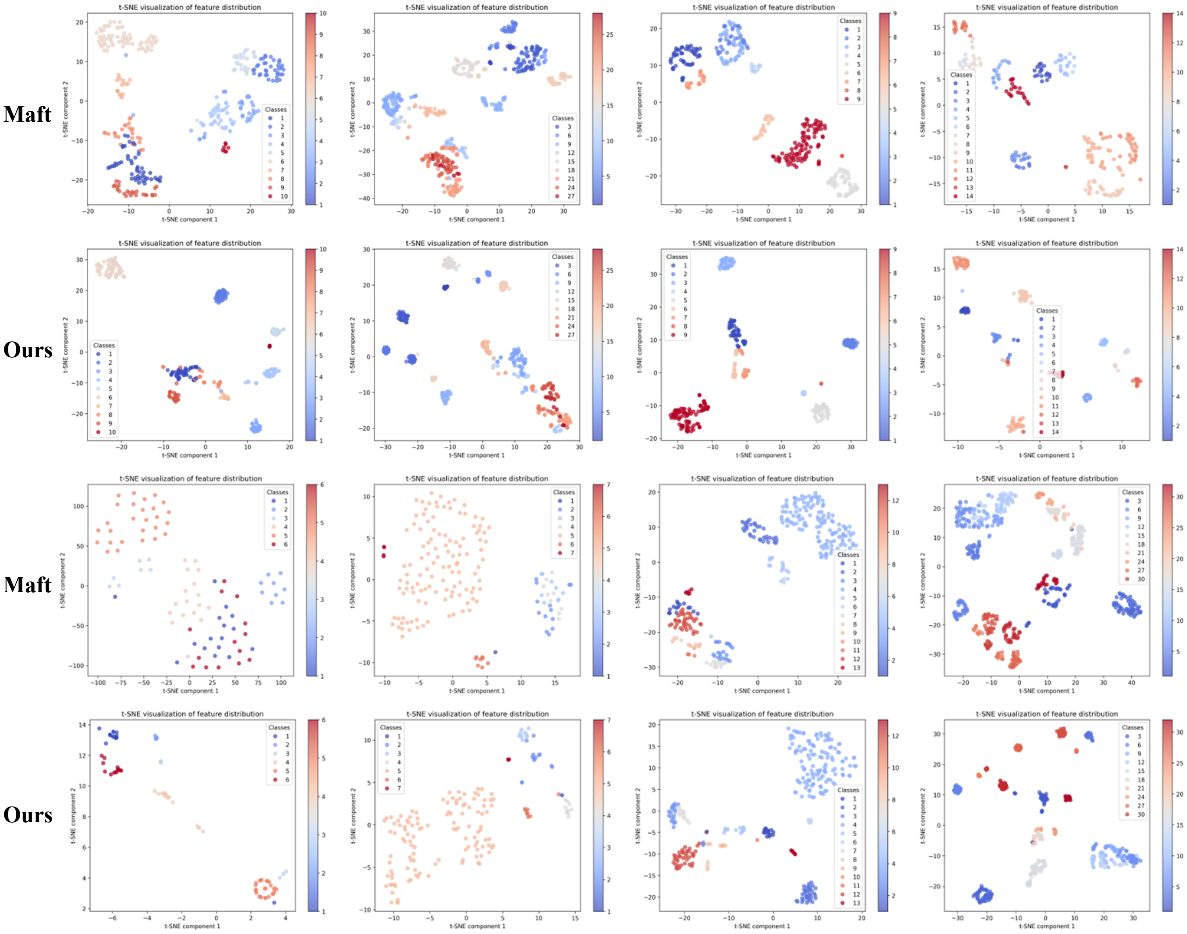

Figure 2. **More T-SNE visualization of the superpoint-level feature distributions on ScanNetV2 validation set.**

2024.

[20] Xiuwei Xu, Huangxing Chen, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Embodiedsam: Online segment any 3d thing in real time. *arXiv preprint arXiv:2408.11811*, 2024.

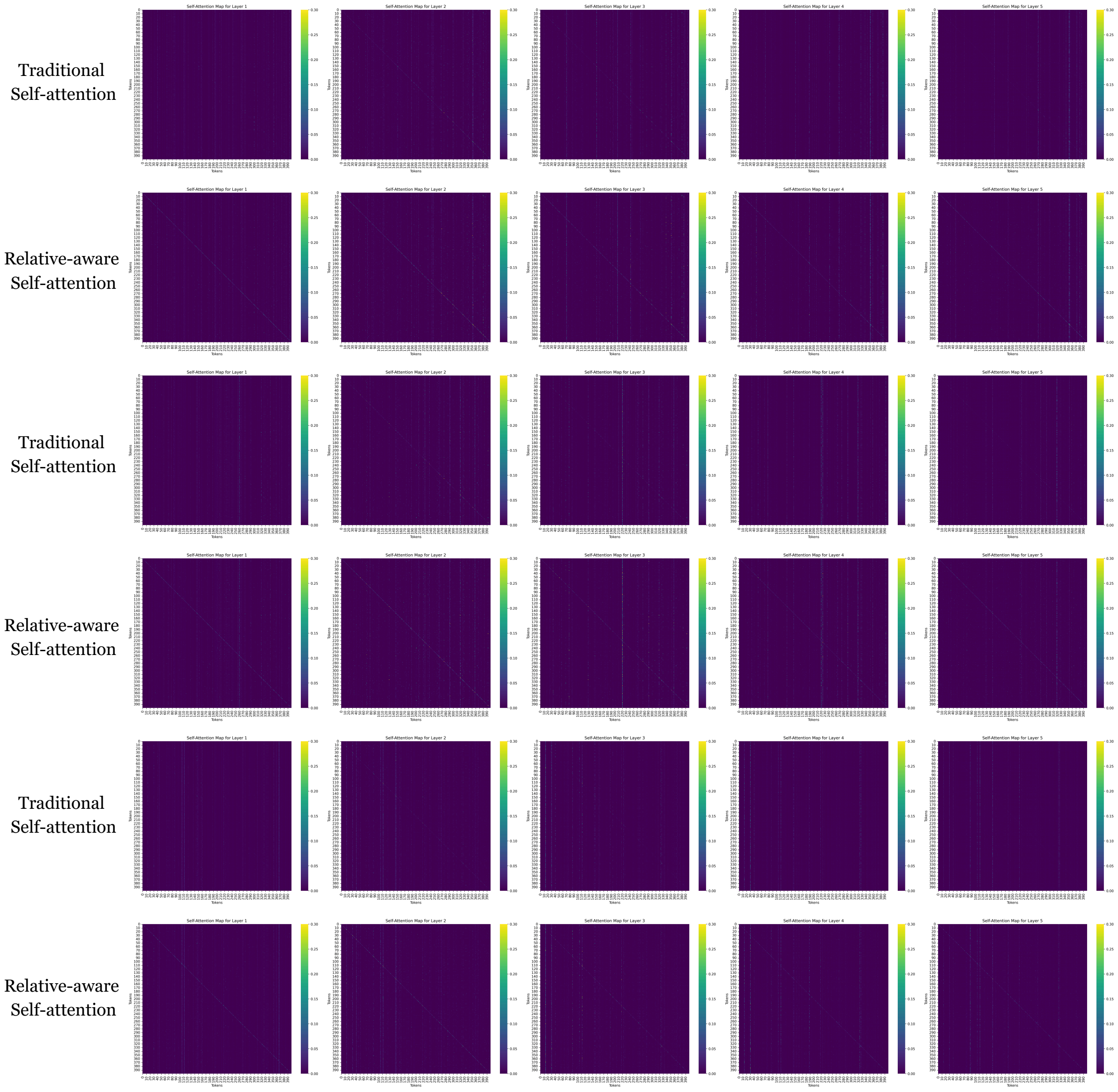

Figure 3. **More comparison of attention maps for traditional self-attention vs. relation-aware self-attention.** We display the progression of attention maps from layer 1 to layer 5.

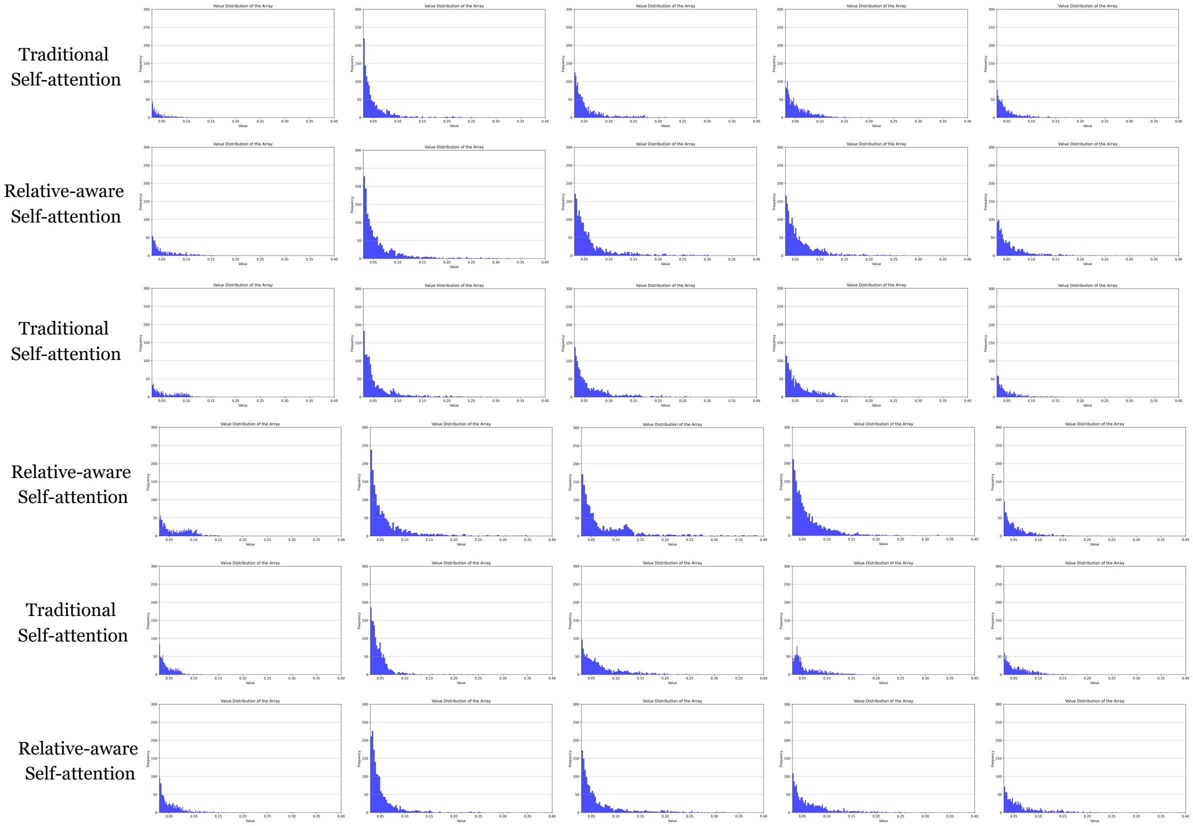

Figure 4. **More comparison of attention weight distributions for traditional self-attention vs. relation-aware self-attention.** The attention weight distributions are also shown from layer 1 to layer 5.

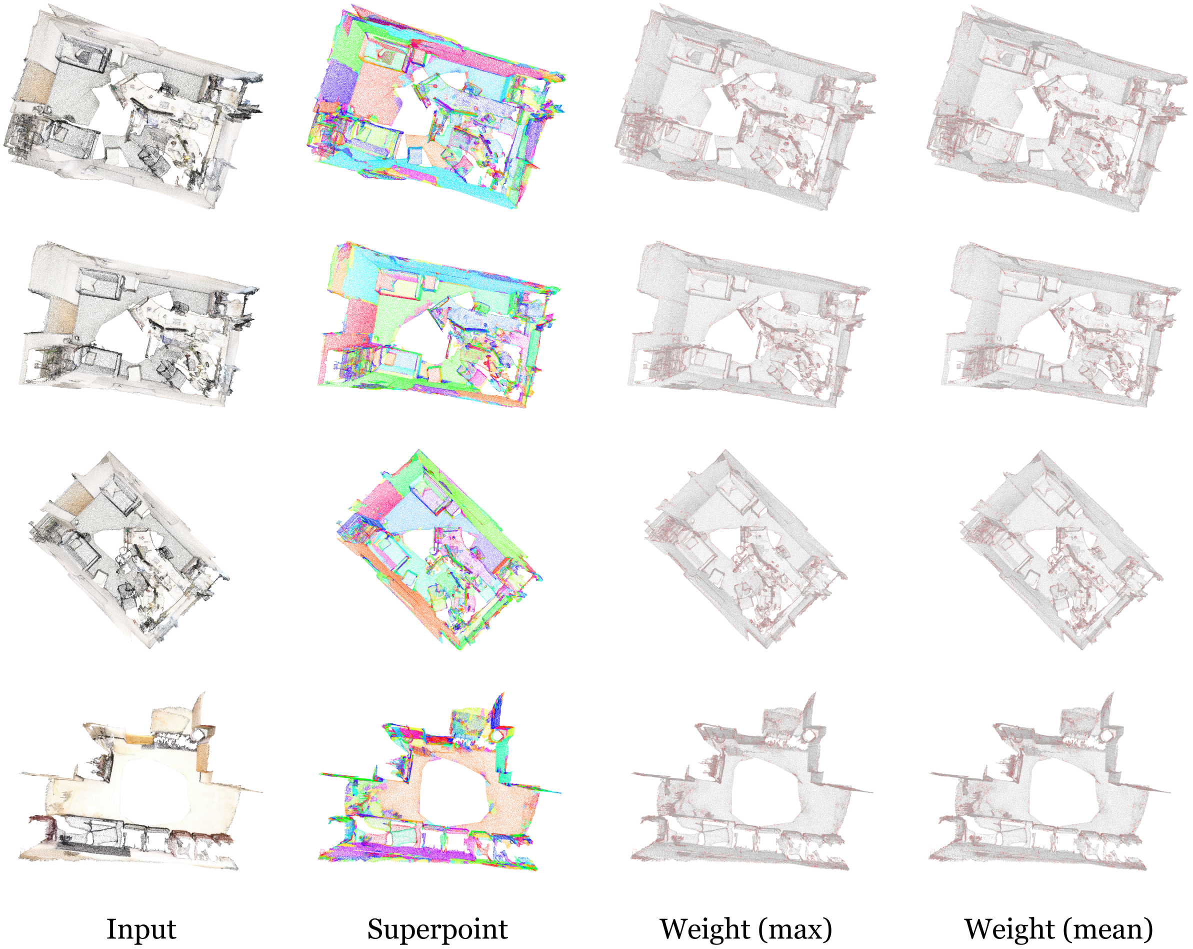

Figure 5. **More visualization of weights in the adaptive superpoint aggregation module.**