Insightful Instance Features for 3D Instance Segmentation

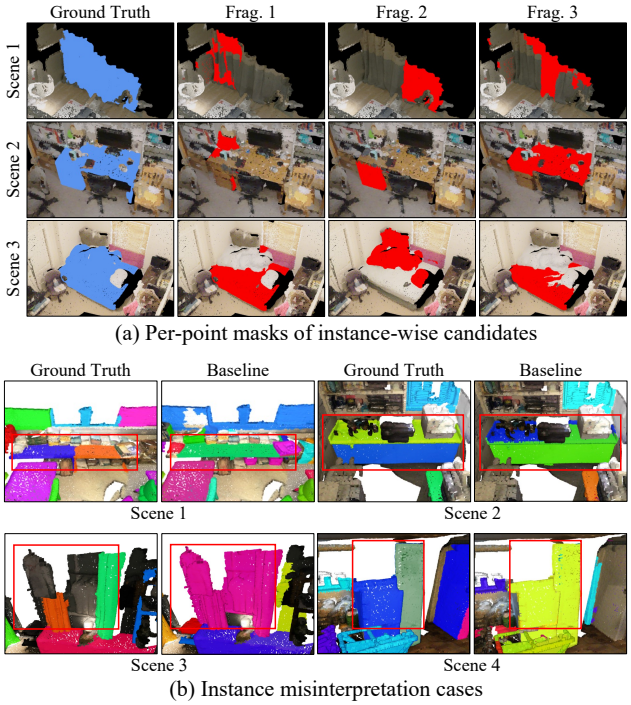
Supplementary Material

Overview

In this supplementary material, we provide further explanations and visualizations of our main paper, "Insightful Instance Features for 3D Instance Segmentation". First, we explore additional challenging cases from previous works (Sec.1). Then, we explain more details about the implementation and large-scale datasets [1, 2, 4, 16] (Sec. 2). Also, we describe our transformer-based architecture (Sec. 3). Moreover, we supply more quantitative and qualitative experimental results to further demonstrate the robustness of our IKNE network for 3D instance segmentation (Sec. 4).

1. Additional Failure Cases

In Fig. 1, we provide additional challenging cases from previous works. (a) First, we empirically observe that their multiple instance candidates usually represent incomplete fragments of the same single instance. (b) Also, they often confuse instances with backgrounds or misunderstand their spatial range. In this work, to address these challenges, we introduce (1) IKA to integrate scattered instance-specific knowledge across multiple instance-wise candidates and (2) ISG to enhance the structural understanding of candidates with essential cues from noise-reduced features.



(a) Per-point masks of instance-wise candidates

(b) Instance misinterpretation cases

Figure 1. Additional examples of two challenging cases from previous works: (a) represents multiple fragments of a single instance, and (b) illustrates misinterpretation of spatial range (Scene 1-2) and instance confusion with backgrounds (Scene 3-4).

2. Experimental Setup

2.1. Datasets

We train and evaluate the overall performance using four landmark datasets for 3D instance segmentation: ScanNetV2 [4], ScanNet200 [16], S3DIS [1], and STPLS3D [2].

ScanNetV2. The ScanNetV2 [4] dataset consists of high-quality, large-scale 3D point data with 1,613 scenes from various room types, such as bedrooms, libraries, and offices. It includes 1,201 training scenes, 312 validation scenes, and 100 hidden test scenes. Each scene is captured using RGB-D cameras and annotated with 20 semantic categories.

ScanNet200. To reflect diverse real-world scenarios, ScanNet200 [16] extends the original ScanNetV2 [4] dataset with fine-grained 200 categories. ScanNet200 enables more practical assessments of how effectively methods can understand rare instances (*e.g.*, *water cooler* or *keyboard piano*) and challenging, long-tail distribution scenes. In our experiments, we evaluate using 18 classes for ScanNetV2 and 198 classes for ScanNet200, excluding *wall* and *floor*.

S3DIS. The S3DIS [1] dataset is large-scale benchmark, comprising a wide range of indoor environments, including 271 scenes from 6 areas within three different buildings. It is categorized with 13 semantic classes, and we utilize all these classes for evaluation. Following the standard protocol [1, 11, 17], we report segmentation scores on Area 5 (Area 5 scenes for evaluation and the others for training) and 6-fold cross-validation (average across all 6 areas).

STPLS3D. The STPLS3D [2] dataset is a extensive aerial photogrammetry dataset containing real and synthetic 3D point clouds. It includes 25 urban scenes covering 6 km², with 14 semantic classes. We use scenes 5, 10, 15, 20, and 25 for evaluation and the rest for training, following [3, 19].

2.2. Implementation Details

In this work, we implement our experimental setup using the PyTorch deep learning framework. For our kernel-based framework, we utilize two point aggregator blocks, each with a ball query radius of 0.2 and 0.4 and 32 neighbors for both layers. We also implement three dynamic convolution layers. We train our model for 120 epochs using a RTX 3090 GPU with a batch size of 12, applying the AdamW optimizer with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}. For the transformer-based pipeline, we utilize a transformer decoder with six layers and eight heads to re-

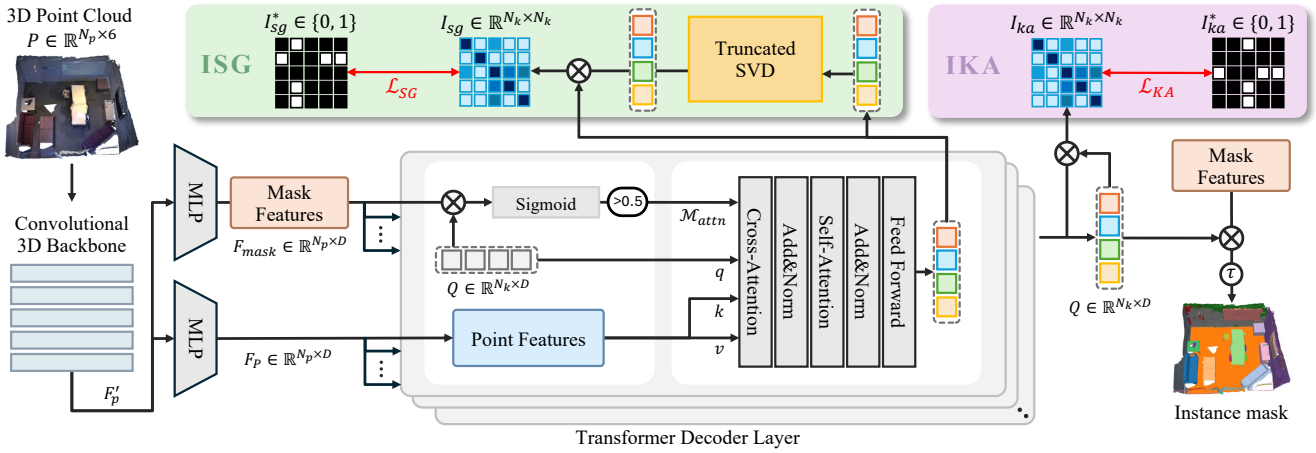


Figure 2. An overview of our transformer-based framework. Built upon the conventional transformer-based architecture, our model consists of four main modules: (1) Sparse Convolutional 3D Backbone; (2) Instance-wise Knowledge Aggregation (IKA); (3) Instance-wise Structural Guidance (ISG); and (4) Mask Transformer Decoder, which iteratively refines instance queries based on attention mechanisms.

fine 400 instance queries. We use Fourier absolute position encoding with a temperature set to 10,000. Also, we train our model for 512 epochs with a batch size of 4, employing the AdamW optimizer with a learning rate of 2×10^{-4} and a weight decay of 5×10^{-2} on a single RTX 3090 GPU.

Regardless of architecture, we set the voxel size to 0.02m for the ScanNet [4] and S3DIS [1] datasets, and 0.3m for the STPLS3D [2] dataset. During training, points are randomly sampled for augmentation with a maximum of 250,000 points, while all points are used for evaluation. This sampling technique is memory-efficient and can also serve as a dropout. In particular, we leverage the superpoint pooling technique [12, 13] on the ScanNet for efficient computation. Also, we set the correlation matrix threshold value τ to 0.9 (exceptionally 0.8 for STPLS3D) for precisely identifying instance candidates representing the same instance, and the top t value as 0.5 for an optimal balance between data compression and information preservation. Moreover, for $\mathcal{L}_{KA}$ and $\mathcal{L}_{SG}$, we set the balance hyperparameters w and $\tilde{w}$ to 5.1×10^{-3} following [9, 20]. Finally, we leverage the Hungarian algorithm [10] for one-to-one matching between the predicted instance masks and the ground truth masks.

3. Our Transformer-based framework

In this section, we explain our instance-wise knowledge enhancement approach based on traditional transformer structure [11, 17]. As shown in Fig. 2, our model consists of four main modules: (1) Sparse Convolutional 3D Backbone; (2) Instance-wise Knowledge Aggregation (IKA), which associates scattered cues of the same single instance; (3) Instance-wise Structural Guidance (ISG), which enhances spatial understandings of instance queries using noise reduced features; and (4) Mask Transformer Decoder, which refines hundreds of candidate queries to contain instance-specific knowledge based on attention mechanisms.

Transformer-based 3DIS Architecture. As in the kernel-based architecture (Sec.3 of our main paper), the sparse convolutional U-Net backbone [7] first takes a colored point cloud $P \in \mathbb{R}^{N_p \times 6}$ as input and extracts full-resolution feature maps $F'_p \in \mathbb{R}^{N_p \times D}$. We then produce F'_p into mask features $F_{mask} \in \mathbb{R}^{N_p \times D}$ and point features $F_p \in \mathbb{R}^{N_p \times D}$ via MLP layers. Following [14, 17], we set zero-initialized non-parametric instance queries $Q \in \mathbb{R}^{N_k \times D}$, referring to point positions sampled with *Farthest Point Sampling* (FPS) [5]. Given the F_{mask}, F_p and Q, the transformer decoder layer iteratively enhances the queries Q through attention layers. Specifically, we employ the masked cross-attention using an intermediate foreground mask $\mathcal{M}_{attn}$. We compute the similarity between Q and F_{mask} using dot product operation, then calculate the probability of the instance mask using the sigmoid function as follows:

$$\mathcal{M}_{attn} = \{m_{i,j} = [\sigma(F_{mask} \cdot Q^T)_{i,j} > 0.5]\} \quad (1)$$

where the threshold value is 0.5 for binary attention mask. With $\mathcal{M}_{attn}$, Q attends to point features F_p in the cross-attention layer to contain instance-specific information as:

$$Q = \text{softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{D} + \mathcal{M}_{attn})\mathbf{V} \quad (2)$$

where $\mathbf{K}$ and $\mathbf{V}$ are linearly projected from F_p, and $\mathbf{Q}$ are from Q. Subsequently, we utilize the standard self-attention layer. Here, the queries, keys, and values are all linear projections of Q. After passing through these layers, we predict the final instance masks using queries from the last layer.

Our Approach. In transformer-based architecture, decoder layers iteratively attend point features, which often contain inherent fuzzy noises due to the sparse and incomplete nature of point clouds. Thus, repetitive layers can lead to noise accumulation in the candidate features during the attention

operations, potentially resulting in spatial range misinterpretations. To tackle this challenge, we introduce the **ISG** network, which regularizes the correlations between the original and clarity-enhanced query features within decoder layers based on a simple yet effective truncated SVD technique [8], as detailed in Sec.3.3 (main paper). Then, the iterative layers continuously enrich instance candidate queries with structural cues. Also, we implement the **IKA** network, which is designed to integrate scattered clues across query features representing the same single instance. The IKA optimizes correlations among instance candidate queries, as outlined in Sec.3.2 (main paper). Ultimately, our model predicts more precise instance masks with highly informative instance query features through our two novel modules.

4. Additional Experiments

4.1. Effectiveness of the IKA

To further validate the effectiveness of our IKA network, we investigate the average variance and standard deviation of instance candidate features across both kernel and transformer based architectures of baselines [11, 15] and those with IKA in Tab. 1 and Fig. 3. We first identify candidates representing the same instance using ground-truth instance masks to ensure fair and more precise comparisons between predicted instance masks from each model. We then calculate the variance and standard deviation of features corresponding to identical instances. Compared to baselines, incorporating IKA consistently achieves lower variance and standard deviation, regardless of the architecture. These results verify that our instance-wise aggregation approach effectively enhances the correlations between candidates from the same instance, establishing meaningful associations.

Method	Avg Variance	Avg StDev
ISBNet (Kernel-based) [15]	2.8441	1.6152
ISBNet w/ IKA	**2.4620**	**1.4199**
MAFT (Transformer) [11]	2.3613	1.5352
MAFT w/ IKA	**2.1278**	**1.2743**

Table 1. Average variance and standard deviation of instance features for kernel / transformer based models and those with IKA.

4.2. Effectiveness of the Our 3DIS framework

We provide t-SNE [18] visualizations of instance candidate features clustered using the density-based spatial clustering (DBSCAN) [6] algorithm to further qualitatively demonstrate the significance of the our framework. As shown in Fig. 4, the candidate features in the baseline method [15] are wildly scattered without patterns in the feature space, resulting in multiple fragments. In contrast, our method produces relatively distinctive clusters for the same scene, with clusters that are accurate to the number of instances. These

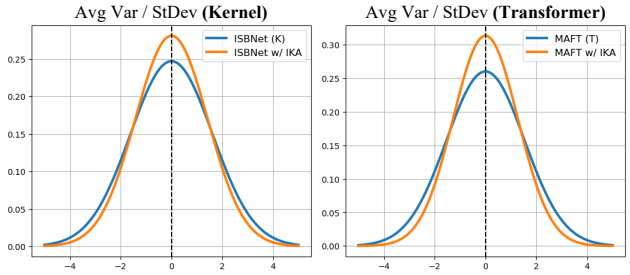


Figure 3. Distributions of the average variance and standard deviation of instance features across kernel and transformer based baselines (blue) and those with the IKA (orange) network.

qualitative findings confirm that our IKA and ISG networks effectively handle hundreds of instance candidate features.

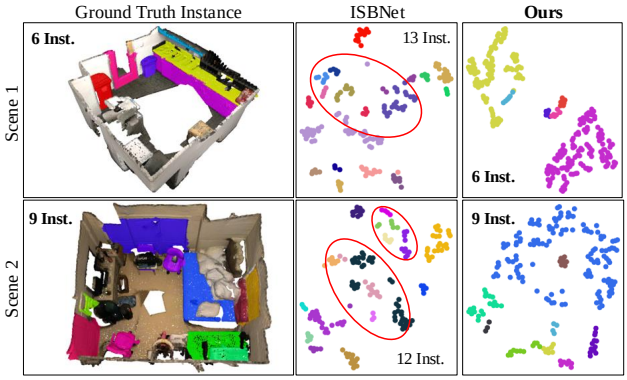


Figure 4. t-SNE [18] visualizations of instance candidate features from kernel-based baseline (ISBNet [15]) and our IKNE method.

4.3. Threshold τ for the STPLS3D Dataset

The threshold τ starts from insights on cosine similarity, which ranges from $[-1, 1]$, where -1 for opposite, 0 for unrelated, and 1 for similar vectors. In our case, we normalize it to $[0, 1]$ (Sec. 3.2, main); thus, 0.5 represents unrelated. We then determine the optimal τ over 0.5 via experiments in Tab.8 of our main paper. In addition to analyzing τ on the ScanNetV2 [4] and S3DIS [1] datasets, we further evaluate a range of τ values on the STPLS3D [2] dataset to identify the threshold that best facilitates robust candidate selection for the same instance. As shown in Tab. 2, the model effectively specifies instance candidates likely to represent the same instance with a somewhat lower threshold of 0.8 for the STPLS3D, compared to 0.9 for both ScanNetV2 and S3DIS. This difference is probably because scenes from the STPLS3D dataset include relatively monotonous large instances, such as buildings and vehicles, unlike ScanNetV2 or S3DIS, which contain more intricate indoor props.

4.4. Correlation Regularization Terms.

Inspired by the self-supervised mechanisms [20], we softly guide highly correlated instance-wise features to be closer

Architecture	Threshold τ	mAP	mAP_{50}
Baseline (Transformer) [17]	-	63.4	79.2
Ours	0.6	61.8	78.8
Ours	0.7	64.5	80.8
Ours	0.8	**64.9**	**81.2**
Ours	0.9	64.7	81.0

Table 2. Ablation study to investigate the correlation threshold τ for instance-wise candidate identification on the STPLS3D [2].

to each other in the latent space, enhancing their solidarity (Eq.6 and Eq.10). Specifically, we utilize dynamically generated pseudo-binary labels to regularize correlations. This element-wise regularizing strategy is conceptually comparable to standard cross-entropy loss, which measures the difference between predicted probabilities and ground-truth distributions. Therefore, to compare the two loss functions, we replace our regularization loss ($\mathcal{L}_{KA}$ and $\mathcal{L}_{SG}$) with standard cross-entropy loss. Here, we set all other settings constant. In Tab. 3, both approaches outperform the baseline [15]; however, ours with cross-entropy loss yields lower performance than our original method. We consider that this performance gap comes from the differences in how each loss handles negative pairs. Cross-entropy loss aims to minimize the disparity between predictions and true labels without explicitly addressing negative pairs. On the other hand, ours considers both positive and negative pairs, reducing the position-wise distances between predicted and target matrices. This strategy encourages the model to establish valuable connections among highly correlated candidates, while minimizing confusion from irrelevant knowledge of unrelated candidates. In conclusion, these results validate the effectiveness of our regularization approach.

Method	ScanNet Val	S3DIS Area 5
	mAP / mAP_{50}	mAP / mAP_{50}
Baseline [15]	56.8 / 73.3	56.3 / 67.5
Ours w/ Cross Entropy	61.2 / 79.7	58.6 / 70.4
Ours	**62.9 / 81.8**	**61.1 / 73.0**

Table 3. Ablation study to compare our correlation regularization terms ($\mathcal{L}_{KA}$ and $\mathcal{L}_{SG}$) with traditional cross-entropy loss.

4.5. Discussion on the Number of Instances

We investigate the number of instances within scenes from various datasets, including ScanNetV2 [4], S3DIS [1], and STPlS3D [2], in Tab. 4. We randomly sampled around 30% of scenes from each dataset and computed the minimum, maximum, and average number of instances. On average, the S3DIS, which consists of a wide range of indoor environments such as exhibition and educational spaces, includes more instances (34.5) per scene than the other two datasets. The ScanNetV2, which contains rooms of various sizes, from small bathrooms to large conference rooms, has

Dataset	min Inst.	max Inst.	avg Inst.
ScanNetV2 [4]	3	104	15.2
S3DIS [1]	6	90	34.5
STPLS3D [2]	2	93	25.2

Table 4. Minimum (min), maximum (max), and average (avg) number of instances per scene from various datasets.

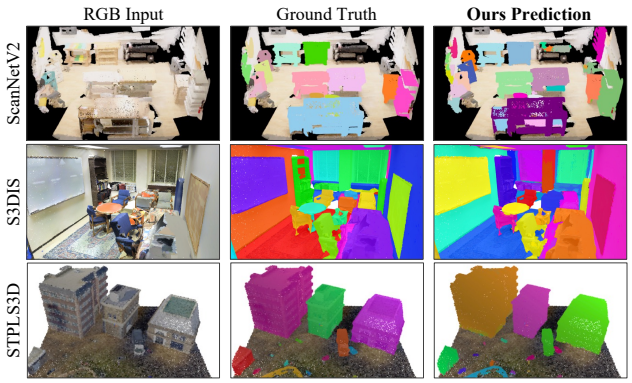


Figure 5. Global view visualizations of predicted instance masks with ground truth from scenes containing many instances across various datasets, ScanNetV2 [4], S3DIS [1], and STPLS3D [2].

a relatively lower average instance count (15.2) per scene but occasionally includes a maximum of 104 instances. In Fig. 5, we also visualize examples of scenes, especially including a large number of instances. Since our approach regularizes features based on correlations among all hundreds of instance candidates, it remains effective regardless of the total instance numbers. While it may be less effective in rare cases where the number of objects exceeds the number of candidates, such scenarios are highly uncommon.

4.6. Visual Comparison

In this section, we present additional qualitative visualization results of our framework, compared to existing state-of-the-art models: ISBNet (kernel-based, K) [15] and MAFT (transformer-based, T) [11], in Fig. 6 and Fig. 7. We visualize the predicted semantic (Sem.) and instance (Inst.) results with corresponding ground truth on the ScanNetV2 [4] dataset, using red colored boxes to highlight the critical differences for better comparison. First, as shown in Fig. 6, our method outperforms existing methods in precisely classifying a single instance into one category without fragments. In particular, compared to baseline models, ours more accurately identifies large instances like a *sofa* (Scene 4) or *cabinet* (Scene 5). Moreover, as shown in Fig. 7, ours distinguishes objects clearer; for example, ours precisely captures their spatial range in Scenes 12-14, where objects are closely adjacent. These results underscore the effectiveness of our novel modules, which enhance instance-wise knowledge for understanding complex real-world environments.

Figure 6. Qualitative comparisons of 3DIS performance on the ScanNetV2 [4] dataset. We visualize semantic (Sem.) masks of ISBNet (kernel-based, K) [15], MAFT (transformer-based, T) [11] and ours based on both architectures with Ground Truth (GT) masks. The key differences are highlighted using red-colored boxes for better comparison. Note that the color map (top right) represents semantic labels.

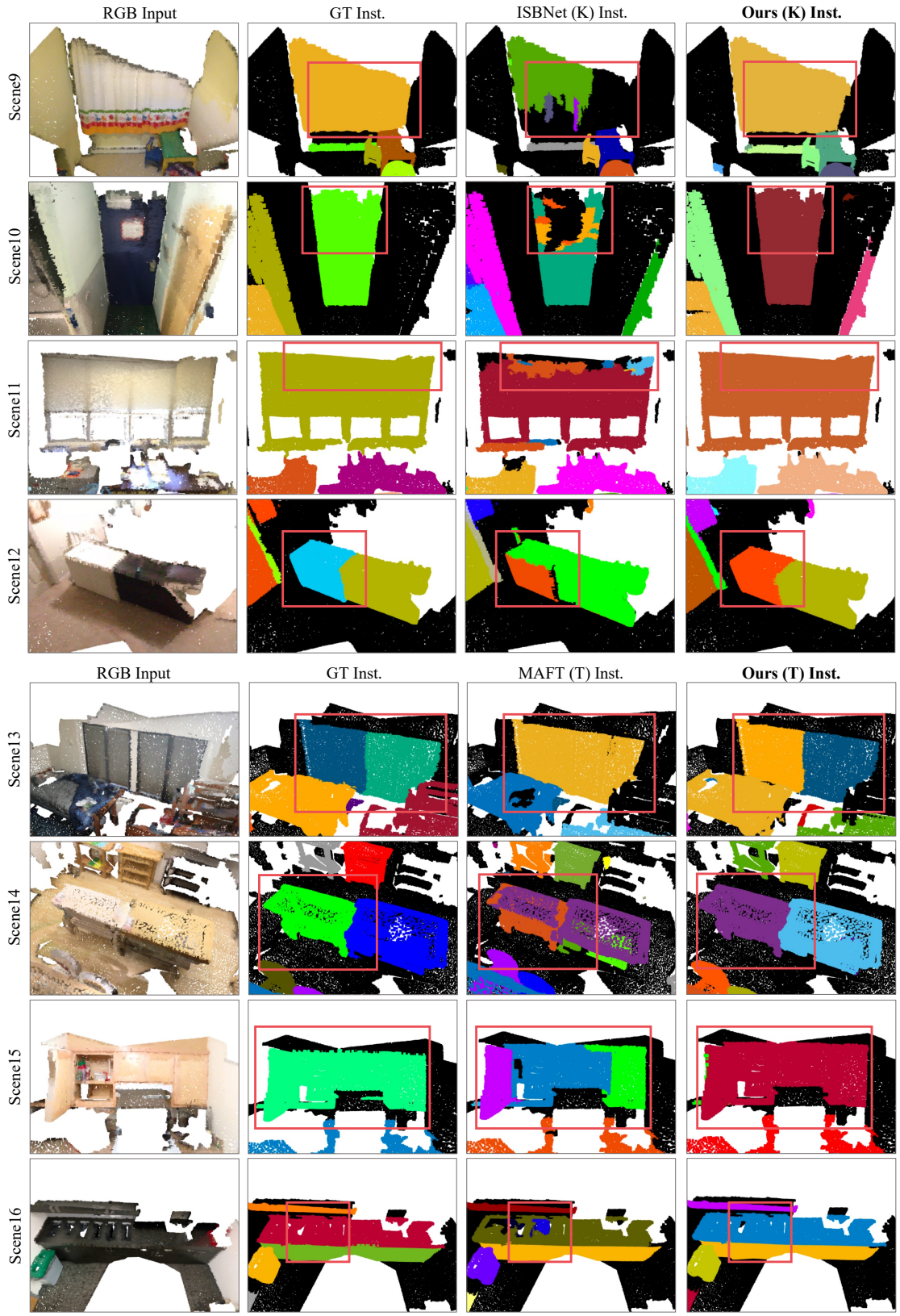


Figure 7. Qualitative comparisons of 3DIS performance on the ScanNetV2 [4] dataset. We visualize instance (Inst.) masks of ISBNet (kernel-based, K) [15], MAFT (transformer-based, T) [11] and ours based on both architectures with Ground Truth (GT) masks. The key differences are highlighted using red-colored boxes for better comparison. Best viewed in color.

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 1, 2, 3, 4

[2] Meida Chen, Qingyong Hu, Zifan Yu, Hugues Thomas, Andrew Feng, Yu Hou, Kyle McCullough, Fengbo Ren, and Lucio Soibelman. Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. *arXiv preprint arXiv:2203.09065*, 2022. 1, 2, 3, 4

[3] 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. 1

[4] 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. 1, 2, 3, 4, 5, 6

[5] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. The farthest point strategy for progressive image sampling. *IEEE transactions on image processing*, 6(9):1305–1315, 1997. 2

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 3

[7] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 2

[8] Per Christian Hansen. The truncated svd as a method for regularization. *BIT Numerical Mathematics*, 27:534–553, 1987. 3

[9] Sujin Jang, Dae Ung Jo, Sung Ju Hwang, Dongwook Lee, and Daehyun Ji. Stxd: Structural and temporal cross-modal distillation for multi-view 3d object detection. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[10] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 2

[11] 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. 1, 2, 3, 4, 5, 6

[12] Loic Landrieu and Mohamed Boussaha. Point cloud oversegmentation with graph-structured deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7440–7449, 2019. 2

[13] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4558–4567, 2018. 2

[14] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021. 2

[15] 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. 3, 4, 5, 6

[16] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, pages 125–141. Springer, 2022. 1

[17] 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. 1, 2, 4

[18] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 3

[19] 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. 1

[20] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021. 2, 3