

MSTA3D: Multi-scale Twin-attention for 3D Instance Segmentation

Anonymous Authors

1 QUALITATIVE RESULTS

In this supplementary material, we provide further qualitative results to validate the effectiveness of the proposed method compared to existing methods including ISBNNet [3], SPFormer [5], and MAFT [2].

ScanNetV2 [1]. As demonstrated in the top examples of Figure 1, there are instances of over-segmentation observed in the chair outputs generated by the SPFormer [5] and MAFT [2] methods. Furthermore, in the bottom figure, both SPFormer [5] and ISBNNet [3] produce over-segmented instances with the class "wall" (background). This indicates that incorporating spatial constraints, the proposed box regularizer, improves segmentation accuracy, especially for nearby objects sharing the same label. Moreover, the integration of low-scale superpoints with high-scale superpoints helps remove background noise, thus reducing over-segmentation in the results.

Similarly, in the top examples of Figure 2, it is observed that the class "desk" is predicted as multiple distinct objects in the outputs of ISBNNet [3] and SPFormer [5]. This indicates that the sampling algorithm used in ISBNNet [3] is limited to large objects due to its sampling radius. In addition, relying solely on high-scale superpoints, as in the approach of SPFormer [5], proves ineffective for identifying large objects when distant superpoints lack tight connections. Over-segmentation also occurs in the "wall" (background) class for all three methods, ISBNNet [3], SPFormer [5], and MAFT [2]. In the bottom examples of Figure 2, all compared methods exhibit over-segmentation in the "wall" (background) and "sink" classes, whereas the proposed method alleviates this issue, demonstrating the reliability of our model's mask predictions supported by bounding box information.

ScanNet200 [4]. The over-segmentation issue is also obvious in the ScanNet200 benchmark [4]. In Figure 3, we compare our results with those of ISBNNet [3]. In the top examples of Figure 3, ISBNNet [3] show over-segmentation in the "door" class with interference from background points. Similarly, in the bottom examples of Figure 3, ISBNNet [3] encounters similar problems with the "curtain" and "piano" classes due to the inflexible sampling radius in its algorithm. This highlights the advantage of using multi-scale approaches, making the proposed model more adaptable in detecting objects of various sizes and thereby significantly improving performance.

REFERENCES

- [1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5828–5839.
- [2] Xin Lai, Yuhui Yuan, Ruihang Chu, Yukang Chen, Han Hu, and Jiaya Jia. 2023. Mask-attention-free transformer for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3693–3703.
- [3] Tuan Duc Ngo, Binh-Son Hua, and Khoi Nguyen. 2023. ISBNNet: 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*. 13550–13559.

- [4] David Rozenberszki, Or Litany, and Angela Dai. 2022. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*. Springer, 125–141.
- [5] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. 2023. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 2393–2401.

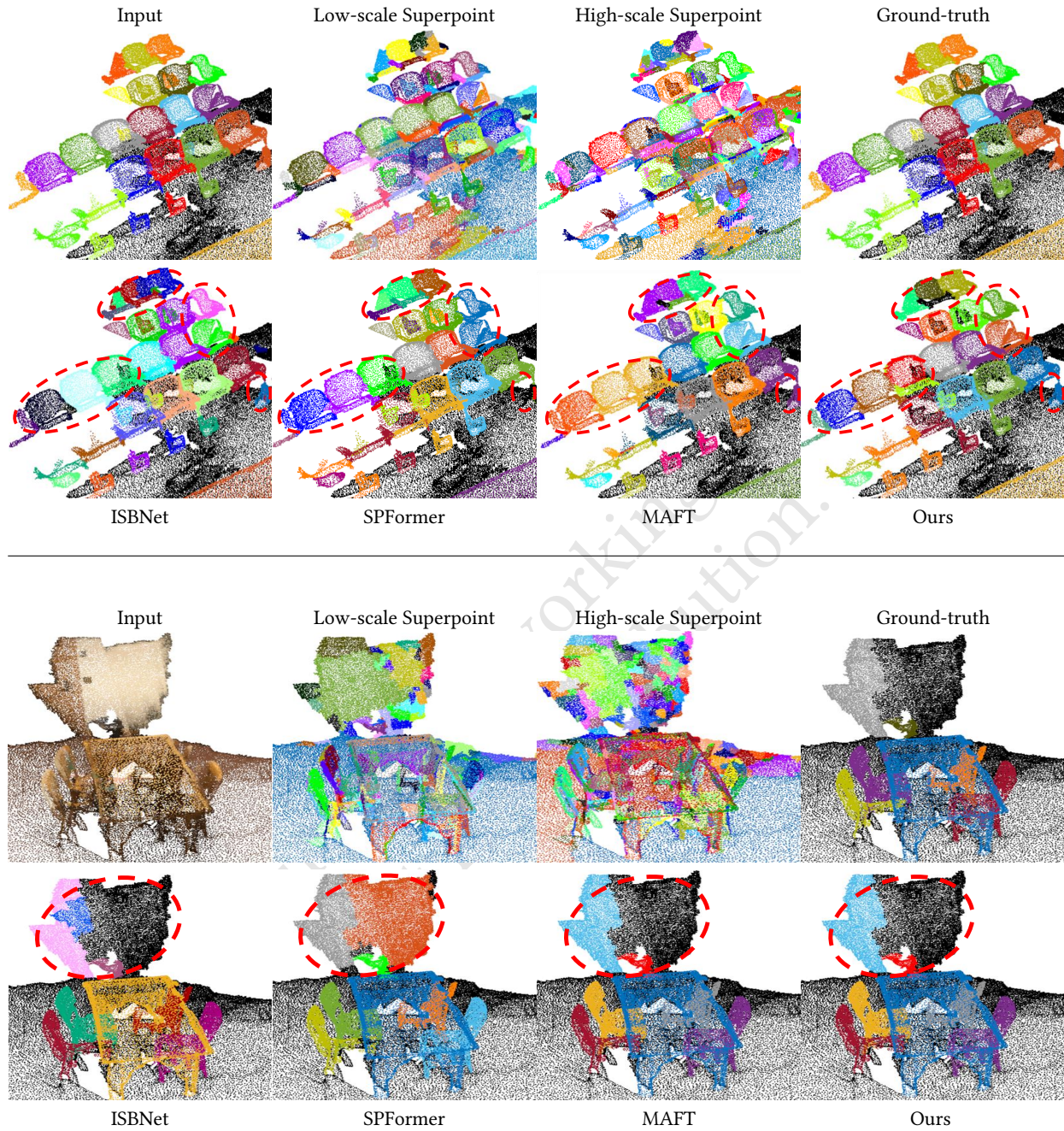


Figure 1: Qualitative comparison of the proposed model with other methods on ScanNetV2. The proposed method mitigates inaccurate instance prediction by introducing a spatial regularizer that integrates scene-wise features and instance-specific features.

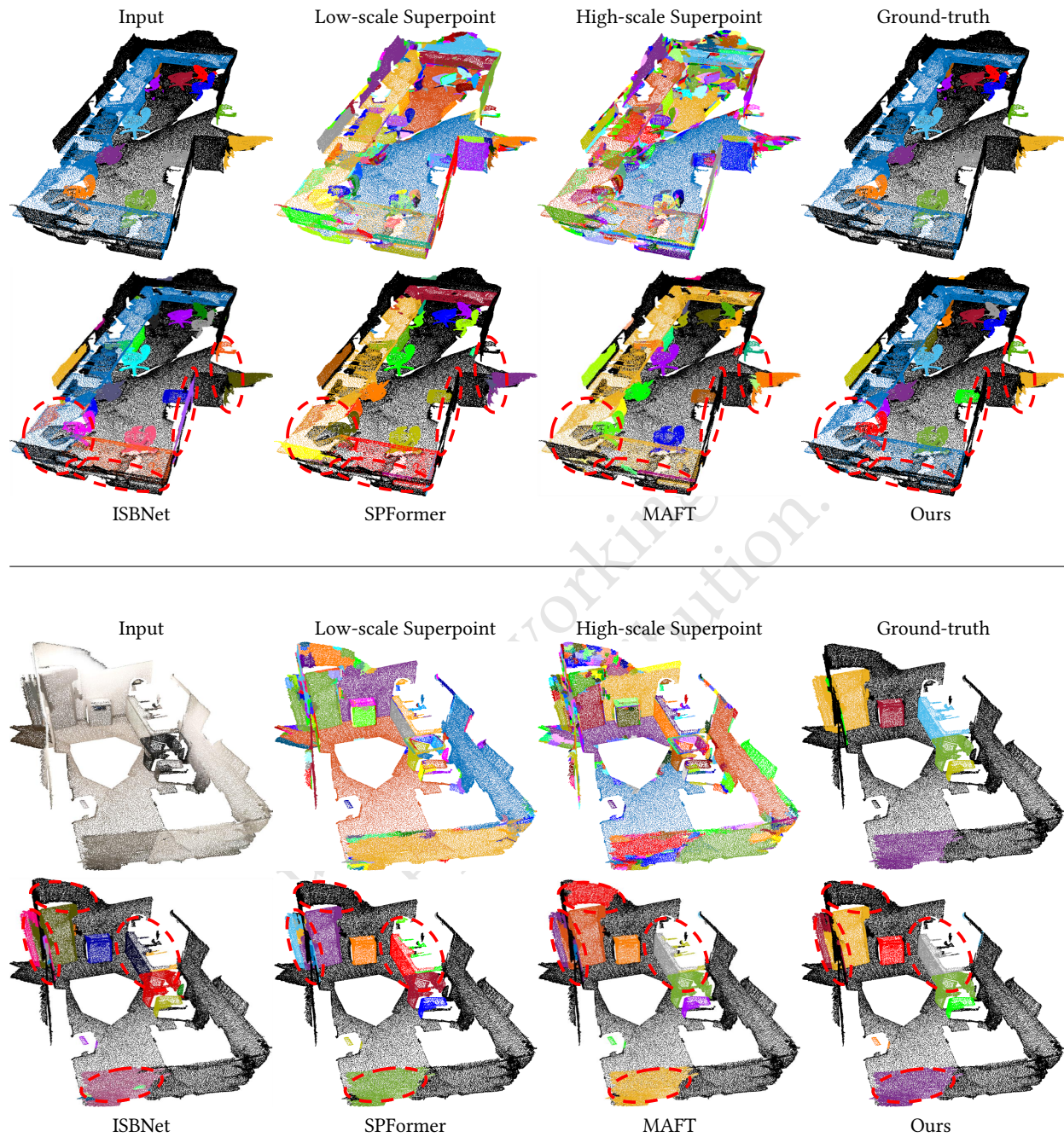


Figure 2: Qualitative comparison of the proposed model with other methods on ScanNetV2. The proposed method overcomes over-segmentation problems by utilizing multi-scale feature representation and spatial constraints.

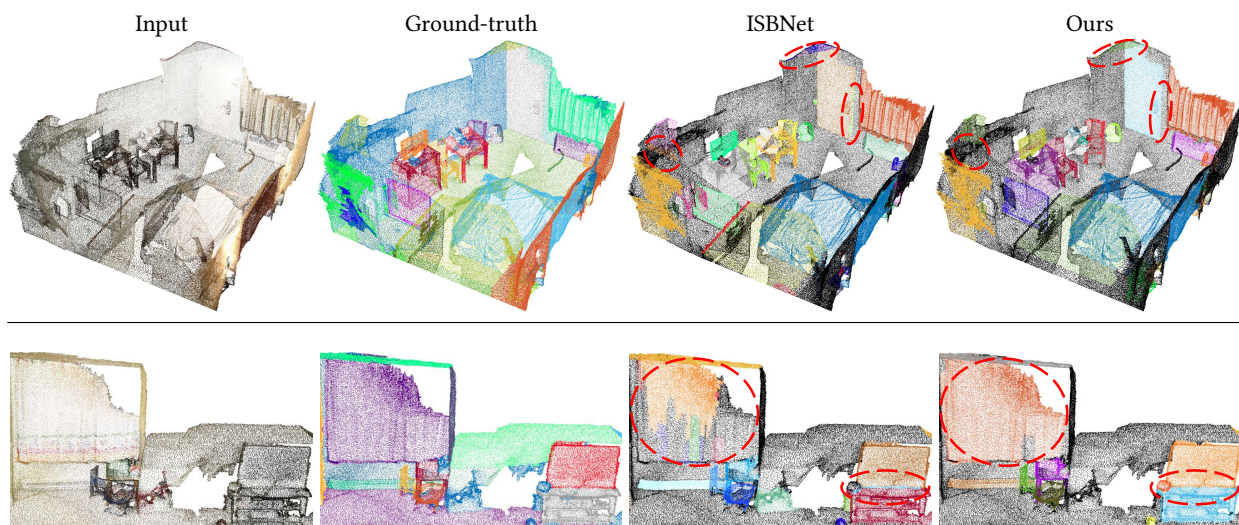


Figure 3: Qualitative comparison of the proposed model with ISBNet on ScanNet200. The proposed method overcomes over-segmentation problems by utilizing multi-scale feature representation and spatial constraints.