## REFERENCES


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*, pp. 1534–1543, 2016.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 213–229. Springer, 2020.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 357–366, 2021a.

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*, pp. 15467–15476, 2021b.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022.

Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3075–3084, 2019.

Spconv Contributors. Spconv: Spatially sparse convolution library. `https://github.com/traveller59/spconv`, 2022.

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*, pp. 5828–5839, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2849–2858, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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*, pp. 9031–9040, 2020.

Maciej Halber, Yifei Shi, Kai Xu, and Thomas Funkhouser. Rescan: Inductive instance segmentation for indoor rgbd scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2541–2550, 2019.

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*, pp. 2940–2949, 2020.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

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*, pp. 354–363, 2021.

Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4421–4430, 2019.

Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2989–2998, 2023.

Haiyong Jiang, Feilong Yan, Jianfei Cai, Jianmin Zheng, and Jun Xiao. End-to-end 3d point cloud instance segmentation without detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12796–12805, 2020a.

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*, pp. 4867–4876, 2020b.

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

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*, pp. 9256–9266, 2019.

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*, pp. 3693–3703, 2023.

Ville V Lehtola, Harri Kaartinen, Andreas Nüchter, Risto Kaijaluoto, Antero Kukko, Paula Litkey, Eija Honkavaara, Tomi Rosnell, Matti T Vaaja, Juho-Pekka Virtanen, et al. Comparison of the selected state-of-the-art 3d indoor scanning and point cloud generation methods. *Remote sensing*, 9(8):796, 2017.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3041–3050, 2023.

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*, pp. 2783–2792, 2021.

Shih-Hung Liu, Shang-Yi Yu, Shao-Chi Wu, Hwann-Tzong Chen, and Tyng-Luh Liu. Learning gaussian instance segmentation in point clouds. *arXiv preprint arXiv:2007.09860*, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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*, pp. 18516–18526, 2023.

Jiahao Lu, Jiacheng Deng, and Tianzhu Zhang. Bsnet: Box-supervised simulation-assisted mean teacher for 3d instance segmentation. *arXiv preprint arXiv:2403.15019*, 2024.

Alessandro Manni, Damiano Oriti, Andrea Sanna, Francesco De Pace, and Federico Manuri. Snap2cad: 3d indoor environment reconstruction for ar/vr applications using a smartphone device. *Computers & Graphics*, 100:116–124, 2021.

Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, volume 3, pp. 850–855. IEEE, 2006.

Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Towards end-to-end lane detection: an instance segmentation approach. In *2018 IEEE intelligent vehicles symposium (IV)*, pp. 286–291. IEEE, 2018.

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*, pp. 13550–13559, 2023.

Kyeong-Beom Park, Minseok Kim, Sung Ho Choi, and Jae Yeol Lee. Deep learning-based smart task assistance in wearable augmented reality. *Robotics and Computer-Integrated Manufacturing*, 63:101887, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

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

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.

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*, pp. 4060–4069, 2024.

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*, volume 37, pp. 2393–2401, 2023.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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*, pp. 2708–2717, 2022.

Chuxin Wang, Jiacheng Deng, Jianfeng He, Tianzhu Zhang, Zhe Zhang, and Yongdong Zhang. Long-short range adaptive transformer with dynamic sampling for 3d object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2569–2578, 2018.

Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. Associatively segmenting instances and semantics in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4096–4105, 2019.

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*, pp. 235–252. Springer, 2022.

Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. *Advances in neural information processing systems*, 32, 2019.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3947–3956, 2019.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6881–6890, 2021.

Min Zhong, Xinghao Chen, Xiaokang Chen, Gang Zeng, and Yunhe Wang. Maskgroup: Hierarchical point grouping and masking for 3d instance segmentation. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2022.

# A Appendix

You may include other additional sections here.

## A.1 Overview

This supplementary material provides more model and experimental details to understand our proposed method. After that, we present more experiments to demonstrate the effectiveness of our methods. Finally, we show a rich visualization of our modules.

## A.2 More Model Details

**Sparse UNet.** For ScanNetV2 Dai et al. (2017), ScanNet200 Rozenberszki et al. (2022), and ScanNet++ Yeshwanth et al. (2023), we employ a 5-layer U-Net as the backbone, with the initial channel set to 32. Unless otherwise specified, we utilize coordinates, colors, and normals as input features. Our method incorporates 6 layers of Transformer decoders, with the head number set to 8, and the hidden and feed-forward dimensions set to 256 and 1024, respectively. For S3DIS Armeni et al. (2016), following Mask3D Schult et al. (2022), we utilize Res16UNet34C Choy et al. (2019) as the backbone and employ 4 decoders to attend to the coarsest four scales. This process is repeated 3 times with shared parameters. The dimensions for the decoder's hidden layer and feed-forward layer are set to 128 and 1024, respectively.

**Transformer Decoder Layer.** In this layer, we use superpoint-level features $F_{\text{sup}}$ and their corresponding positions $P_{\text{sup}}$ as key and value, with content queries $Q^c$ and position queries $Q^p$ as query. The specific network architecture can be seen in Figure 6, which is identical to Maft's Lai et al. (2023) transformer decoder layer. Therefore, more relevant equations and details can be directly referred to Maft's main text.

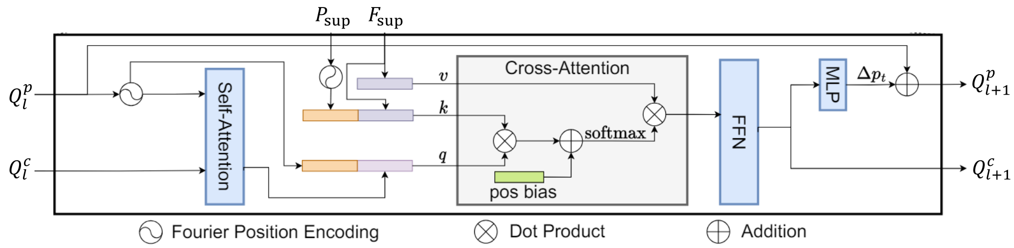



Figure 6: **The architecture of the transformer decoder layer.** The figure is taken from the main text of Maft.

**Matching and Loss.** Existing methods depend on semantic predictions and binary masks for matching queries with ground truths. Building upon Maft Lai et al. (2023), our approach integrates center distance into Hungarian Matching Kuhn (1955). To achieve this, we modify the formulation of matching costs as follows:

$$\mathcal{C}_{cls}(p, \overline{p}) = CE(CLASS_p, CLASS_{\overline{p}}), \tag{8}$$

$$\mathcal{C}_{dice}(p, \overline{p}) = DICE(MASK_p, MASK_{\overline{p}}), \tag{9}$$

$$\mathcal{C}_{bce}(p, \overline{p}) = BCE(MASK_p, MASK_{\overline{p}}), \tag{10}$$

$$\mathcal{C}_{center}(p, \overline{p}) = L1(Center_p, Center_{\overline{p}}), \tag{11}$$

$$\mathcal{C}(p, \overline{p}) = \lambda_{cls}\mathcal{C}_{cls}(p, \overline{p}) + \lambda_{dice}\mathcal{C}_{dice}(p, \overline{p}) + \lambda_{bce}\mathcal{C}_{bce}(p, \overline{p}) + \lambda_{center}\mathcal{C}_{center}(p, \overline{p}), \tag{12}$$

where $p$ and $\overline{p}$ denotes a predicted and ground-truth instance, $\mathcal{C}$ represents the matching cost matrix, and $\lambda_{cls}, \lambda_{dice}, \lambda_{bce}, \lambda_{center}$ are the hyperparameters. Here, $\lambda_{cls}, \lambda_{dice}, \lambda_{bce}, \lambda_{center}$ are the same as $\lambda_1, \lambda_2, \lambda_3, \lambda_4$. Next, we perform Hungarian Matching on $\mathcal{C}$, and then supervise the Hungarian Matching results according to Equation 7

**Non-Maximum Suppression.** Non-maximum suppression (NMS) is a common post-processing operation used in instance segmentation. In fact, for some previous methods, applying NMS to the final layer predictions has consistently led to performance improvements, as shown in Table 12. However, if we apply NMS to the concatenated outputs, as described in Section 1 lines 63-65, a significant decrease in performance occur. The specific reasons for this performance decrease are twofold. Firstly, NMS heavily relies on confidence scores, retaining only the masks with the highest confidence among the duplicates. However, these confidence scores are often inaccurate, leading to the retention of masks that are not necessarily of the best quality. Since the concatenated outputs contain a large number of duplicate masks (almost every mask has duplicates), this results in a significant reduction in performance. Secondly, NMS requires manual selection of a threshold. If the threshold is set too high, it cannot effectively filter out duplicate masks; if it is set too low, it tends to discard useful masks. The more complex the output, the more challenging it becomes to select an optimal threshold. Therefore, for concatenated outputs, it is difficult to find an optimal threshold for effective filtering.

| Method | mAP | AP@50 | AP@25 |
|---|---|---|---|
| SPFormer | 56.7 | 74.8 | 82.9 |
| SPFormer+NMS | 57.2 | 75.9 | 83.5 |
| SPFormer+COE | 55.7 | 73.4 | 81.8 |
| Maft | 58.4 | 75.2 | 83.5 |
| Maft+NMS | 59.0 | 76.1 | 84.3 |
| SPFormer+COE | 57.3 | 73.5 | 81.8 |
| Ours | 61.1 | 78.2 | 85.6 |
| Ours+NMS | **61.7** | **79.5** | **86.5** |

Table 12: **The effectiveness of the NMS.** COE refers to concatenating the outputs of each layer and then conducting NMS.

### A.3 MORE DISCUSSION

**Details on achieving a strong correlation.** The positions of sampling points in Mask3D are not related to the positions of the corresponding predicted instances. In fact, this lack of correlation results in the query's lack of interpretability, we cannot clearly understand why this query predicts this object, thus hindering intuitive optimization. Both QueryFormer and Maft address this by adding a $\mathcal{C}_{center}$ term when calculating the Hungarian matching cost matrix, which represents the distance between the query coordinates and the ground truth instance center. Additionally, they update the query coordinates layer by layer, making the matched query progressively closer to the GT instance center. With this design, the position of the query becomes correlated with the position of the corresponding predicted instance, facilitating intuitive improvements in the distribution of query initialization by QueryFormer and Maft (Query Refinement Module and Learnable Position Query).

**Detail classification on Hierarchical Query Fusion Decoder.** We aim to give poorly updated queries a new opportunity for updating. It is important to note that this is a copy operation, so we retain both pre-updated and post-updated queries, thus not "limiting the transformer decoder in its ability to swap objects." This approach provides certain queries with an opportunity for entirely new feature updates and offers more diverse matching options during Hungarian matching. This re-updating and diverse selection mechanism clearly enhances recall rates because our design implicitly includes a mechanism: for instances that are difficult to predict or poorly predicted, if the updates are particularly inadequate, the corresponding queries will be retained and accumulated into the final predictions. For example, if a query $Q_i^3$ from the third layer is updated in the fourth layer to become $Q_i^4$ and experiences a significant deviation, the network will retain $Q_i^3$ and pass both $Q_i^3$ and $Q_i^4$ to the fifth layer. After being updated in the fifth layer, $Q_i^3$ becomes $\hat{Q_i^3}$. If $\hat{Q_i^3}$ does not significantly differ from $Q_i^3$, the model will not retain $Q_i^3$ further and will only pass $\hat{Q_i^3}$ to the sixth layer. If $\hat{Q_i^3}$ shows a significant difference from $Q_i^3$ , the model will continue to retain $Q_i^3$. Through this process, teh model can continuously retain the queries that are poorly updated, accumulating them into the final prediction.

### A.4 MORE IMPLEMENTATION DETAILS

On ScanNet200 Rozenberszki et al. (2022), we train our model on a single RTX3090 with a batch size of 8 for 512 epochs. We employ AdamW Loshchilov & Hutter (2017) as the optimizer and PolyLR as the scheduler, with a maximum learning rate of 0.0002. Point clouds are voxelized with a size of 0.02m. For hyperparameters, we tune $\mathcal{S}, L, K, \mathcal{D}_1, \mathcal{D}_2$ as 500, 500, 3, 40, 3 respectively. $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ in Equation 7 are set as 0.5, 1, 1, 0.5, 0.5. On ScanNet++ Yeshwanth et al. (2023), we train our model on a single RTX3090 with a batch size of 4 for 512 epochs. The other settings are the same as ScanNet200. On S3DIS Armeni et al. (2016), we train our model on a single A6000 with a batch size of 4 for 512 epochs and adopt onecycle scheduler. For hyperparameters, we tune $\mathcal{S}, L, K, \mathcal{D}_1, \mathcal{D}_2$ as 400, 400, 3, 40, 3 respectively. $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ in Equation 7 are set as 2, 5, 1, 0.5, 0.5.

### A.5 DETAILED RESULTS

The detailed results for each category on ScanNetV2 validation set are reported in Table 13. As the table illustrates, our method achieves the best performance in 16 out of 18 categories. The detailed results for certain categories on ScanNet++ test set are presented in Table 17. As indicated by the table, the significant performance improvement highlights the effectiveness of our method in managing denser point cloud scenes across a broader range of categories.

### A.6 MORE ABLATION STUDIES

**Difference in Recall and AP across different decoder layers.** As depicted in Table 18, we conduct an ablation study on ScanNetV2 validation set to examine the impact of our proposed HQFD on recall and AP. From the table, it is evident that the recall of Maft decreases at the fifth layer, consequently leading to a decline in the corresponding AP and influencing the final prediction results. In contrast, our approach, which incorporates HQFD, ensures a steady improvement in recall, thereby

| Method | mAP | bathtub | bed | bookshe. | cabinet | chair | counter | curtain | desk | door | other | picture | frige | s. curtain | sink | sofa | table | toilet | window |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SoftGroup Vu et al. (2022) | 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 Wu et al. (2022) | 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 Schult et al. (2022) | 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 |
| QueryFormer Lu et al. (2023) | 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 Lai et al. (2023) | 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 | **61.7** | **83.5** | **62.3** | **48.1** | **50.6** | **84.1** | **45.0** | **57.4** | 42.1 | **57.3** | **61.8** | **67.8** | **59.9** | **68.8** | **61.1** | **55.3** | **66.6** | **95.3** | 42.6 |

Table 13: **Full quantitative results of mAP on ScanNetV2 validation set.** Best performance is in boldface.

| Method | mAP | bathtub | bed | bookshe. | cabinet | chair | counter | curtain | desk | door | other | picture | frige | s. curtain | sink | sofa | table | toilet | window |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PointGroup Jiang et al. (2020b) | 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 Zhong et al. (2022) | 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 Han et al. (2020) | 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 Chen et al. (2021b) | 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 Liang et al. (2021) | 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 |
| DKNet Wu et al. (2022) | 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 |
| SPFormer Sun et al. (2023) | 54.9 | 74.5 | 64.0 | 48.4 | 39.5 | 73.9 | **31.1** | 56.6 | 33.5 | 46.8 | 49.2 | 55.5 | 47.8 | 74.7 | 43.6 | 71.2 | 54.0 | 89.3 | 34.3 |
| Maft Lai et al. (2023) | 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 | **60.6** | **92.6** | 70.2 | 51.5 | **50.2** | 73.2 | 28.2 | **59.8** | 38.6 | 48.9 | **54.2** | **63.5** | **71.6** | 75.1 | 47.6 | **74.3** | **58.7** | **95.8** | 36.0 |

Table 14: **Full quantitative results of mAP on the ScanNetV2 test set. Best performance is in boldface.**

| Method | AP@50 | bathtub | bed | bookshe. | cabinet | chair | counter | curtain | desk | door | other | picture | frige | s. curtain | sink | sofa | table | toilet | window |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PointGroup Jiang et al. (2020b) | 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 Zhong et al. (2022) | 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 Han et al. (2020) | 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 Chen et al. (2021b) | 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 Liang et al. (2021) | 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 |
| DKNet Wu et al. (2022) | 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 |
| SPFormer Sun et al. (2023) | 77.0 | 90.3 | 90.3 | 80.6 | 60.9 | 88.6 | 56.8 | **81.5** | 70.5 | 71.1 | 65.5 | 65.2 | 68.5 | 100.0 | 78.9 | 80.9 | 77.6 | 100.0 | 58.3 |
| Maft Lai et al. (2023) | 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 | **81.0** | **100.0** | **93.4** | 85.4 | **74.3** | 88.9 | **57.5** | 71.4 | **81.0** | 66.9 | **72.9** | 70.7 | **80.9** | **100.0** | 81.4 | **90.2** | **81.4** | **100.0** | 62.5 |

Table 15: **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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PointGroup Jiang et al. (2020b) | 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 Zhong et al. (2022) | 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 Han et al. (2020) | 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 Chen et al. (2021b) | 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 Liang et al. (2021) | 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 |
| DKNet Wu et al. (2022) | 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 |
| SPFormer Sun et al. (2023) | 85.1 | 100.0 | 99.4 | 80.6 | 77.4 | 94.2 | 63.7 | **84.9** | 85.9 | 88.9 | 72.0 | 73.0 | 66.5 | 100.0 | 91.1 | 86.8 | 87.3 | 100.0 | 79.6 |
| Maft Lai et al. (2023) | 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 | **88.2** | **100.0** | 97.9 | 88.2 | **87.9** | 93.7 | 70.3 | 74.9 | **91.5** | 87.5 | **79.5** | 74.0 | **82.0** | **100.0** | **99.4** | **92.3** | **89.1** | **100.0** | 78.8 |

Table 16: **Full quantitative results of AP@25 on the ScanNetV2 test set. Best performance is in boldface.**

guaranteeing a consistent enhancement in AP. This favorable effect on the final output results is attributed to the design of this moudle.

| | |
|---|---|
| Submission creation date | 5 Aug, 2024 |
| Last edited | 5 Aug, 2024 |

3D semantic instance results

Metric: AP

| Info | avg ap | bathtub | bed | bookshelf | cabinet | chair | counter | curtain | desk | door | otherfurniture | picture | refrigerator | shower curtain | sink | sofa | tabl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.606 | 0.926 | 0.702 | 0.515 | 0.502 | 0.732 | 0.282 | 0.598 | 0.386 | 0.489 | 0.542 | 0.635 | 0.716 | 0.751 | 0.476 | 0.743 | 0.58 |

Figure 7: **The mAP result of our method on ScanNetV2 test set.**

| | |
|---|---|
| Submission creation date | 5 Aug, 2024 |
| Last edited | 5 Aug, 2024 |

3D semantic instance results

Metric: AP 50%

| Info | avg ap 50% | bathtub | bed | bookshelf | cabinet | chair | counter | curtain | desk | door | otherfurniture | picture | refrigerator | shower curtain | sink | sofa | tabl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.810 | 1.000 | 0.934 | 0.854 | 0.743 | 0.889 | 0.575 | 0.714 | 0.810 | 0.669 | 0.729 | 0.707 | 0.809 | 1.000 | 0.814 | 0.902 | 0.81 |

Figure 8: **The AP@50 result of our method on ScanNetV2 test set.**

| | |
|---|---|
| Submission creation date | 5 Aug, 2024 |
| Last edited | 5 Aug, 2024 |

3D semantic instance results

Metric: AP 25%

| Info | avg ap 25% | bathtub | bed | bookshelf | cabinet | chair | counter | curtain | desk | door | otherfurniture | picture | refrigerator | shower curtain | sink | sofa | tabl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.882 | 1.000 | 0.979 | 0.882 | 0.879 | 0.937 | 0.703 | 0.749 | 0.915 | 0.875 | 0.795 | 0.740 | 0.820 | 1.000 | 0.994 | 0.923 | 0.89 |

Figure 9: **The AP@25 result of our method on ScanNetV2 test set.**

| | |
|---|---|
| Submission creation date | 17 Nov, 2024 |
| Last edited | 17 Nov, 2024 |

3D semantic instance results

Metric: AP | Column Sorting: Column Sort Alphabetically

| Info | avg ap | head ap | common ap | tail ap | alarm clock | armchair | backpack | bag | ball | bar | basket | bathroom cabinet | bathroom counter | bathroom stal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.294 | 0.412 | 0.267 | 0.182 | | | 0.487 | 0.078 | 0.083 | 0.000 | 0.000 | 0.000 | | 0.01: |

Figure 10: **The mAP result of our method on ScanNet200 test set.**

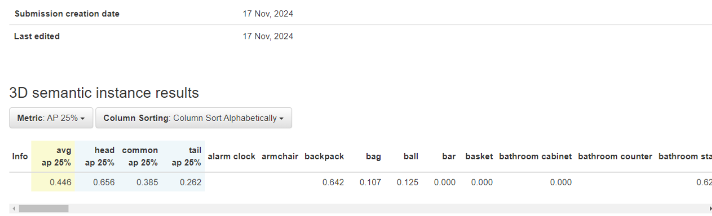
| | |
|---|---|
| Submission creation date | 17 Nov, 2024 |
| Last edited | 17 Nov, 2024 |

3D semantic instance results

Metric: AP 50% | Column Sorting: Column Sort Alphabetically

| Info | avg ap 50% | head ap 50% | common ap 50% | tail ap 50% | alarm clock | armchair | backpack | bag | ball | bar | basket | bathroom cabinet | bathroom counter | bathroom stal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.399 | 0.574 | 0.355 | 0.237 | | | 0.608 | 0.102 | 0.125 | 0.000 | 0.000 | 0.000 | | 0.03 |

Figure 11: **The AP@50 result of our method on ScanNet200 test set.**



| | |
|---|---|
| Submission creation date | 17 Nov, 2024 |
| Last edited | 17 Nov, 2024 |

3D semantic instance results

Metric: AP 25% ▾ Column Sorting: Column Sort Alphabetically ▾

| Info | avg ap 25% | head ap 25% | common ap 25% | tail ap 25% | alarm clock | armchair | backpack | bag | ball | bar | basket | bathroom cabinet | bathroom counter | bathroom stal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.446 | 0.656 | 0.385 | 0.262 | | | 0.642 | 0.107 | 0.125 | 0.000 | 0.000 | 0.000 | | 0.62 |

Figure 12: **The AP@25 result of our method on ScanNet200 test set.**

| Method | mAP | bottle | box | ceiling l. | cup | monitor | office c. | white. e. | tv | white. | telephone | tap | tissue b. | trash c. | window | sofa | pillow | plant | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PointGroup Wu et al. (2022) | 8.9 | 0.8 | 2.1 | 57.3 | 13.2 | 37.8 | 82.8 | 0 | 39.0 | 54.7 | 0 | 0 | 0 | 37.2 | 3.5 | 35.7 | 10.1 | 22.5 | ... |
| HAIS Schult et al. (2022) | 12.1 | 3.4 | 3.8 | 55.9 | 16.8 | 49.5 | **87.1** | 0 | 64.1 | 72.5 | 7.2 | 0 | 0 | 29.5 | 4.0 | 49.0 | 14.9 | 25.0 | ... |
| SoftGroup Vu et al. (2022) | 16.7 | 9.4 | 6.2 | 46.7 | 23.2 | 42.8 | 81.3 | 0 | 67.3 | 71.6 | 10.9 | 14.0 | **2.9** | 32.9 | 8.1 | 46.4 | 17.0 | **60.0** | ... |
| Ours | **22.2** | **13.2** | **12.7** | **63.7** | **38.1** | **69.3** | 86.0 | **38.9** | **90.6** | **86.8** | **26.7** | **20.6** | 2.0 | **60.0** | **9.4** | **63.7** | **45.3** | 52.5 | ... |

Table 17: **Full quantitative results of mAP on ScanNet++ test set.** Best performance is in boldface.

| Layer | Ours | | | | Maft | | | |
|---|---|---|---|---|---|---|---|---|
| | Recall@50 | mAP | AP@50 | AP@25 | Recall@50 | mAP | AP@50 | AP@25 |
| 3 | 87.5 | 59.4 | 76.7 | 84.9 | 85.7 | 56.9 | 73.9 | 82.5 |
| 4 | 87.8 (+) | 59.7 (+) | 77.1 (+) | 85.1 (+) | 86.6 (+) | 58.5 (+) | 75.5 (+) | 83.7 (+) |
| 5 | 87.9 (+) | 59.9 (+) | 77.3 (+) | 85.3 (+) | 85.8 (-) | 58.2 (-) | 75.0 (-) | 83.5 (-) |
| 6 | **88.1** (+) | **60.9** (+) | **78.1** (+) | **85.7** (+) | 86.6 (+) | 59.0 (+) | 76.1 (+) | 84.3 (+) |

Table 18: **Difference in Recall and AP across different decoder layers.** (+) indicates an increase compared to the previous layer, while (-) indicates a decrease compared to the previous layer.

**Ablation study on $\mathcal{D}_1$ and $\mathcal{D}_2$ of the Hierarchical Query Fusion Decoder.** $D_1$ represents the number of new added queries in each layer compared to the previous layer, while $D_2$ indicates the layers where the fusion operation is performed. From the table data, we can see that performance decreases significantly when $D_2$=4 compared to $D_2$=3. As analyzed in lines 334-336 in the main text, the queries in the earlier layers have not aggregated enough instance information. Therefore, if $D_2$=4, it means that the queries in the second layer will also participate in the fusion operation, but these queries have only undergone two rounds of feature aggregation, resulting in inaccurate mask predictions. This can affect the operation of the Hierarchical Query Fusion Decoder (HQFD). To ensure the effectiveness of HQFD, we recommend performing the fusion operation on the last half of the decoder layers. In fact, we follow this approach in other datasets as well.

| $\mathcal{D}_1$ | $\mathcal{D}_2$ | mAP | AP@50 | AP@25 |
|---|---|---|---|---|
| 50 | 2 | 61.4 | 78.9 | 86.1 |
| 50 | 3 | 61.5 | 79.2 | 86.3 |
| 50 | 4 | 61.0 | 78.5 | 85.6 |
| 40 | 3 | **61.7** | **79.5** | **86.5** |
| 60 | 3 | 61.3 | 78.8 | 85.9 |

Table 19: **Ablation study on $\mathcal{D}_1$ and $\mathcal{D}_2$ of the Hierarchical Query Fusion Decoder.**

**The effectiveness of the SG in Equation 5.** As illustrated in Table 20, we performed an ablation study on ScanNetV2 validation set to examine the impact of the SG operation in Equation 5. If we do not utilize SG, $Q_0^p$ remains fixed, which hinders its ability to adaptively learn a distribution suitable for all scenarios, thus impacting the overall performance.

| Setting | mAP | AP@50 |
|---|---|---|
| W SG | 61.4 | 79.0 |
| W/o SG | **61.7** | **79.5** |

Table 20: **The effectiveness of the SG in Equation 5.**

**Ablation Study on the hyperparameters in Equation 7.** We perform the experiment in Table 21. Based on the results, we find that the combination 0.5, 1, 1, 0.5, 0.5 yields the best performance.

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_5$ | mAP |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 0.5 | 0.5 | 61.1 |
| 0.5 | 1 | 1 | 0.5 | 0.5 | **61.7** |
| 1.5 | 1 | 1 | 0.5 | 0.5 | 61.4 |
| 0.5 | 0.5 | 1 | 0.5 | 0.5 | 60.8 |
| 0.5 | 1.5 | 1 | 0.5 | 0.5 | 61.5 |
| 0.5 | 1 | 0.5 | 0.5 | 0.5 | 61.0 |
| 0.5 | 1 | 1.5 | 0.5 | 0.5 | 61.2 |
| 0.5 | 1 | 1 | 1 | 0.5 | 61.0 |
| 0.5 | 1 | 1 | 0.5 | 1 | 61.5 |

Table 21: **Ablation Study on the hyperparameters in Equation 7 on ScanNetV2 validation set.**

### A.7 Assets Availability

The datasets that support the findings of this study are available in the following repositories:

ScanNetV2 Dai et al. (2017) at `http://www.scan-net.org/changelog#scannet-v2-2018-06-11` under the ScanNet Terms of Use. ScanNet200 Rozenberszki et al. (2022) at `https://github.com/ScanNet/ScanNet` under the ScanNet Terms of Use. ScanNet++ Yeshwanth et al. (2023) at `https://kaldir.vc.in.tum.de/scannetpp` under the ScanNet++ Terms of Use. S3DIS Armeni et al. (2016) at `http://buildingparser.stanford.edu/dataset.html` under Apache-2.0 license. The code of our baseline Lai et al. (2023); Sun et al. (2023) is available at `https://github.com/dvlab-research/Mask-Attention-Free-Transformer` and `https://github.com/sunjiahao1999/SPFormer` under MIT license.

### A.8 More Visual Comparison

In Figure 13, we visualize and compare the results of several methods. As shown in this figure's red boxes, our method produces finer segmentation results.

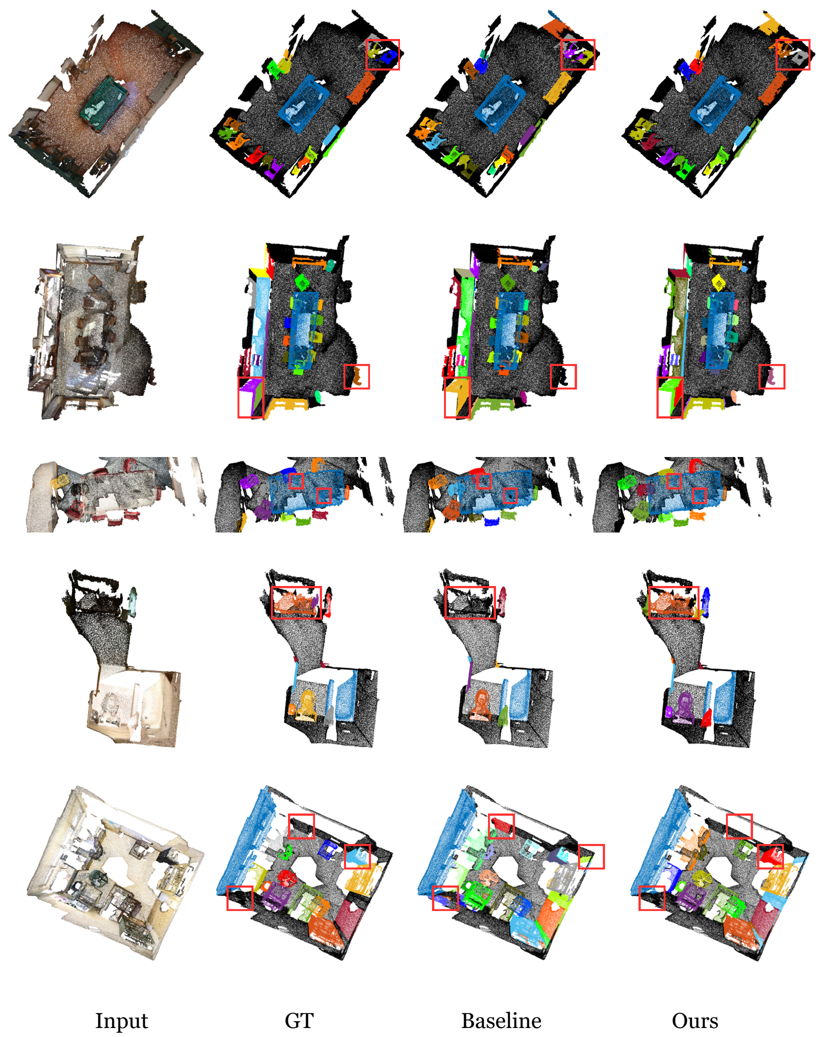



Figure 13: **Additional Visual Comparison on ScanNetV2 validation set.** The red boxes highlight the key regions.

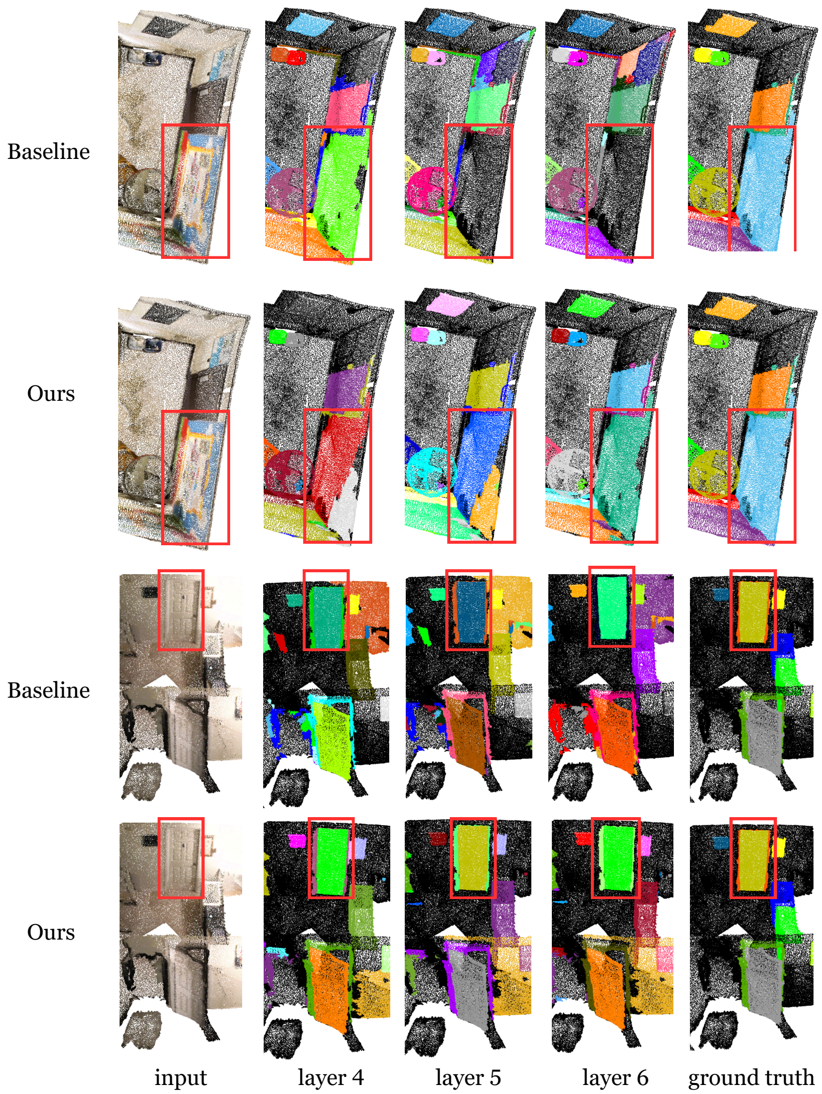



Figure 14: **Visual comparisons between the baseline and our method across different decoder layers on ScanNetV2 validation set.** The red boxes highlight the key regions.

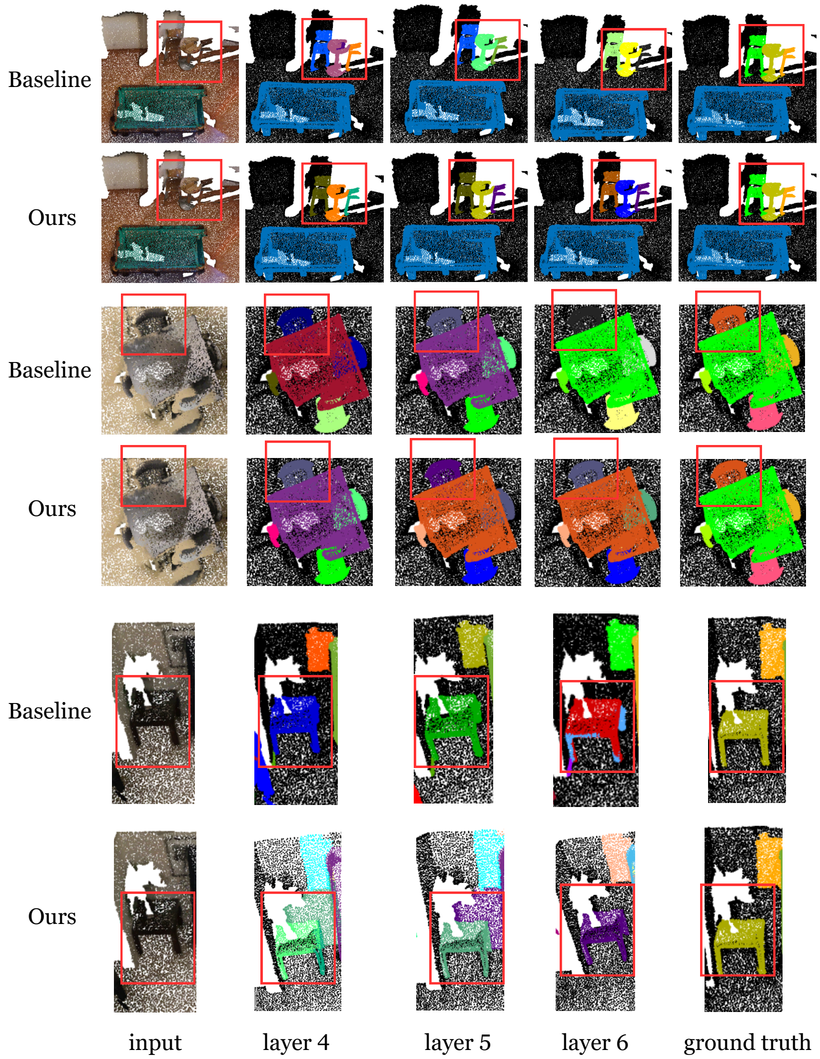



Figure 15: **Visual comparisons between the baseline and our method across different decoder layers on ScanNetV2 validation set.** The red boxes highlight the key regions.

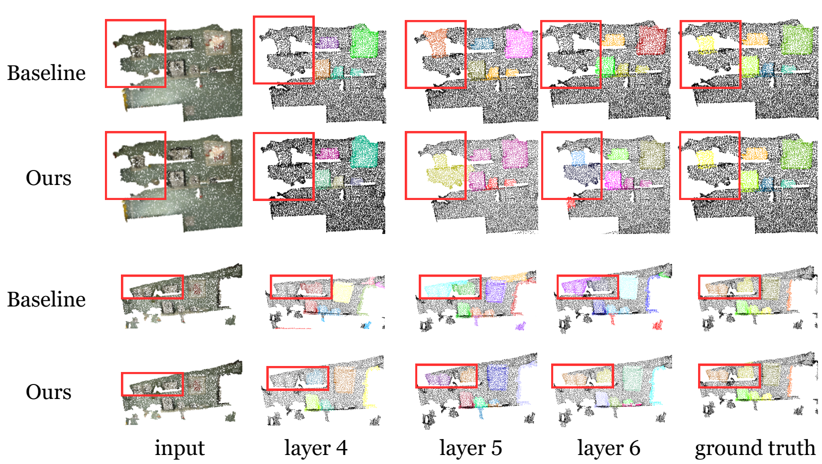



Figure 16: **Visual comparisons between the baseline and our method across different decoder layers on ScanNetV2 validation set.** The red boxes highlight the key regions.