Point Cloud Semantic Segmentation Method Based on Multi-scale Supervision

Zhiyuan Wang [1,*], Shaoqing Liu [1], Wei Jiang [2], Jiaxu Wang [1]

[1] Chinese Flight Test Establishment, Xi'an City, Shaanxi Province 710089, China
[2] Troop 95828, Xi'an City, Shaanxi Province 710089, China

*Corresponding author's e-mail:751491546@qq.com

Abstract. Due to the ability to faithfully reconstruct real-world 3D scenes, point cloud data are widely used in fields such as digital twins and scene understanding. However, manual annotation of large-scale point clouds is inefficient, whereas point-cloud semantic segmentation technology can extract the geometric features of various objects in 3D scenes to generate fine-grained masks for different categories. To simulate a real air combat environment, we constructed safe test conditions on the ground and used LIDAR to scan the fighter jet cockpit. For the precise differentiation of the various avionics systems in the virtual cockpit, we designed a point cloud semantic segmentation network based on multi-scale supervision, which employs an encoder to generate multi-hot labels for supervising feature maps at corresponding scales across multiple decoders. We performed 6-fold cross-validation of our network on the indoor point cloud dataset S3DIS and transferred the model to cockpit point cloud data for segmentation. Ultimately, we achieved an mIoU of 75.3% and an OA of 90.2% on S3DIS.

Keywords: Point cloud semantic segmentation, Multi-scale supervision, Contrastive learning.

1 Introduction

In recent years, the increasingly complex air combat environment has imposed stringent requirements on the upgrading of avionics equipment. Especially during the flight test and evaluation process, the avionics system may be updated once a month, making it difficult for pilots to quickly master the operation methods of the numerous functions of the avionics system. Therefore, digital twins create virtual cockpits by modeling and simulating the aircraft cockpit, which can create a realistic cockpit environment and interact with the complete aircraft simulation system, enabling pilots to practice air tasks on the ground in advance [1]. To rapidly construct a virtual fighter cockpit, we use LiDAR to scan the cockpit and generate 3D point cloud data. However, it is a significant challenge to distinguishing between the complex avionics in a virtual cockpit. Recently, due to the easy accessibility of point clouds, point cloud semantic segmentation technology is widely applied in the field of scene understanding [20, 22, 23, 31], as it enables the extraction and classification of the spatial geometric features of point

K. Othman et al. (eds.), *Proceedings of the 2026 5th International Conference on Mathematical Statistics and Economic Analysis (MSEA 2026)*, Advances in Economics, Business and Management Research 402,
https://doi.org/10.2991/978-94-6239-774-3_33

clouds [2]. Therefore, we utilize the indoor public point cloud dataset S3DIS [3] for training and transfer the feature model to the segmentation of cockpit avionics equipment. In this paper, we use PointNext [4] as the baseline and employ multiple decoders to fuse feature maps of various scales from the encoder. Finally, we utilize contrastive learning [25] to supervise the segmentation mask obtained through upsampling by the decoder [26]. On the indoor semantic segmentation dataset S3DIS, our network achieved a mean Intersection over Union (mIoU) of 90% and an Overall Accuracy (OA) of 90%.

2 METHOD

2.1 Network Architecture

Considering the application environment characterized by limited cockpit space and various of avionics equipment [27], we selected point-based method to train on indoor scene datasets. Furthermore, since the convenience of engineering implementation, we use PointNext as the baseline and optimize it using a multi-decoder architecture and contrastive learning, as shown in Figure 1.

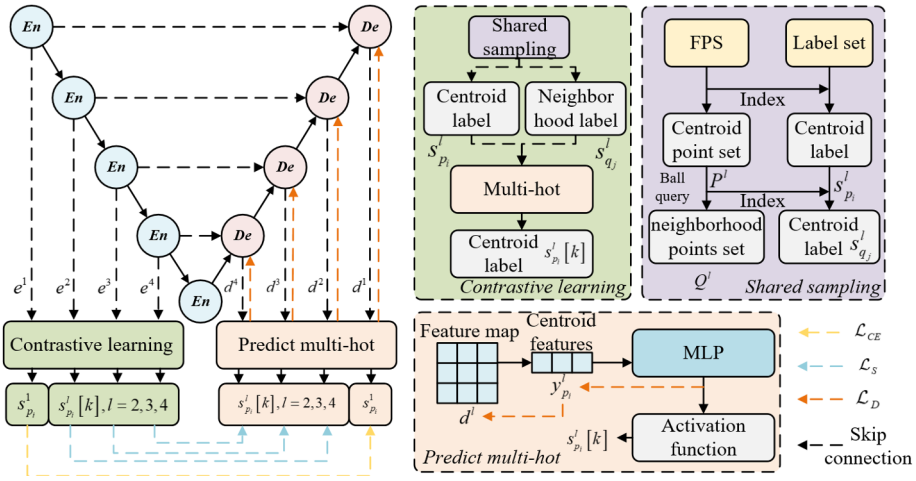

Fig. 1. Network architecture.

Our network consists of a 5-layer encoder and four decoders containing different layers. All decoders transform feature maps of different scales from the encoder into features of a unified scale, producing high-precision segmentation results that incorporate both large-scale contextual semantic information and local detailed features [30]. Furthermore, considering the complex feature interactions and training difficulties associated with various decoders, we employ contrastive learning at each decoder scale to suppress ineffective features generated during upsampling [18]. Specifically, the contrastive learning module combines features from different decoders that share the same scale and predicts multi-class labels, which represent the categories present in the

feature map [28, 29]. Meanwhile, feature maps at various scales within the encoder are extracted to generate ground-truth multi-class labels. Ultimately, the multi-class labels generated by the encoder are used to supervise the multi-class labels predicted by the decoder.

2.2 Contrastive Learning and Multi-scale Supervision

For each encoding layer of the encoder, the sets of centroid points and neighborhood points to be processed are designated as P^l and Q^l, respectively. And the network needs to be trained based on the label information s_p of each point p and classify it into M categories. However, for the intermediate layers in the encoder and decoder, each centroid sampled via farthest point sampling aggregates the features of its K neighboring points. However, the centroid encapsulates all feature representations of the neighboring points. Therefore, it is difficult for contrastive learning to use the intermediate layers of the encoder to provide feature supervision for the intermediate layers of the decoder.

Inspired by [14, 19, 21], we employed multi-hot labels for feature supervision, as shown in Figure 2.

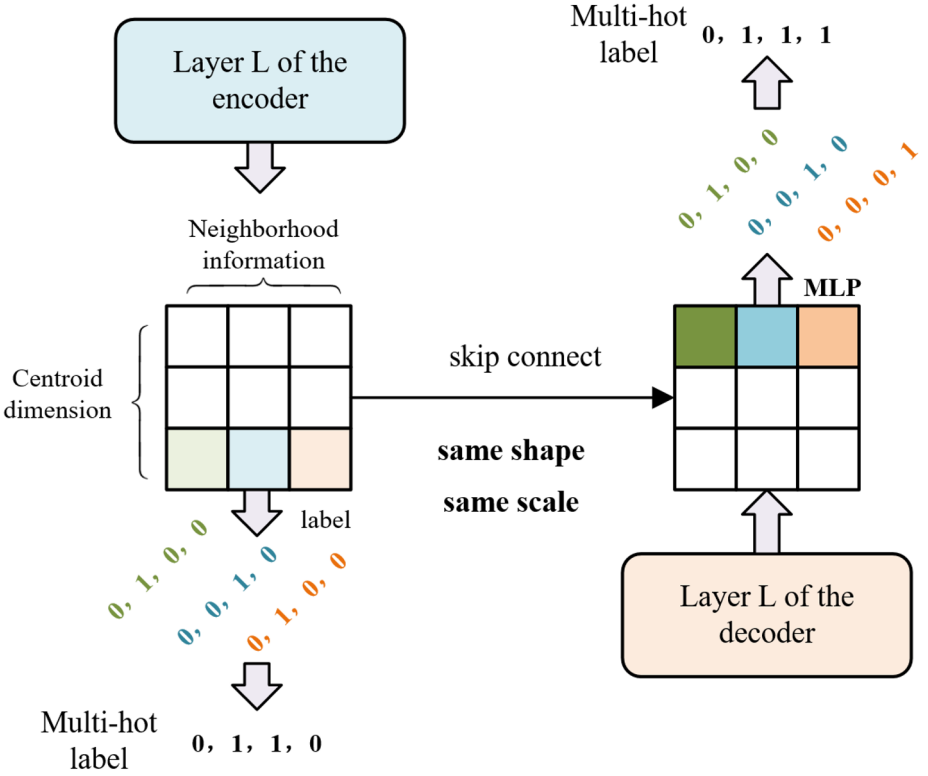

Fig. 2. Generation and alignment of multi-hot labels.

We use feature maps of the same scale from both the encoder and the decoder to generate multi-hot labels, consequently. Therefore, the multi-hot labels generated by the encoder can be used to supervise the multi-hot labels predicted by the decoder at the same scale. Specifically, for P^l and Q^l, the corresponding label information is $s^l_{p_i \in P}$ and $s^l_{q_j \in Q}$, respectively. The multi-class labels of the encoder are shown in (1).

$$s^l_{p_i}[k] = \bigcup_{q_j \in Q^{l-1}} s^{l-1}_{q_j}[k] \tag{1}$$

Where $s^l_{p_i}[k]$ represents the label vector containing the k categories for the centroid point p in layer l, $\cup$ represents the OR operation performed on the one-hot or multi-hot labels of all points in the neighborhood point set.

For decoders without explicit label annotations, we use an MLP to predict the multi-hot labels from the features of each layer. Specifically, a network with an L-layer encoder can employ L-1 decoders to upsample the encoded features from each layer to a unified scale. Set $y^{g,l}_{p_i}$ represents the feature information of the centroid point p_i in the output feature map of the l-th layer of the g-th decoder. And the multi-hot labels $s^{g,l}_{p_i}[k]$ predicted by MLP as shown in (2).

$$s^{g,l}_{p_i}[k] = \sigma\left[MLP\left(y^{g,l}_{p_i}\right)\right], g \in [1, L\text{-}1], l \in [1, g] \tag{2}$$

Where σ represents the activation function.

Therefore, we introduce contrastive learning between encoders and decoders operating at the same scale, as shown in (3) and (4).

$$\mathcal{L}^{g,l}_S = -\frac{1}{MN}\sum_{i=1}^{N}\sum_{k=1}^{M}\mathcal{L}^{g,l}(p_i,k) \tag{3}$$

$$\mathcal{L}^{g,l}(p_i,k) = s^l_{p_i}[k]\log\left(s^{g,l}_{p_i}[k]\right) + \left(1 - s^l_{p_i}[k]\right)\log\left(1 - s^{g,l}_{p_i}[k]\right) \tag{4}$$

Where $\mathcal{L}^{g,l}(p_i,k)$ represents the cross-entropy loss function, which utilizes the multi-hot label $s^l_{p_i}[k]$ of the centroid point p_i in the l-th layer of the encoder to supervise the predicted multi-hot label $s^{g,l}_{p_i}[k]$ of the l-th layer of the g-th encoder. And $\mathcal{L}^{g,l}_S$ averages the loss functions calculated at the centroids of the same scale across all decoders. Finally, the loss functions of all decoders are aggregated to obtain the multi-scale supervision loss function $\mathcal{L}_S$, as shown in (5).

$$\mathcal{L}_S = \frac{1}{(L-1)^2}\sum_{l=1}^{g-1}\sum_{g=1}^{L-1}\mathcal{L}^{g,l}_S \tag{5}$$

However, multiple decoders introduce a large number of intermediate layers containing various inactive negative features. Therefore, for the centroid p_i of the l-th layer in the g-th decoder, we introduce a centrifugal potential function to construct a cross-entropy loss function $\mathcal{L}_D^{g,l}(p_i)$, which amplifies positive features and suppresses negative features [24], as shown in (6) and (7).

$$\Phi(\overline{y_{p_i}^{g,l}}) = -\log\frac{1}{1+e^{-|\overline{y_{p_i}^{g,l}}|}}, \overline{y_{p_i}^{g,l}} = MLP\left(y_{p_i}^{g,l}\right) \tag{6}$$

$$\mathcal{L}_D^{g,l}(p_i) = -\left[u\left(\overline{y_{p_i}^{g,l}}\right)\log\left(\frac{1}{1+e^{-\overline{y_{p_i}^{g,l}}}}\right) + \left(1-u\left(\overline{y_{p_i}^{g,l}}\right)\right)\log\left(\frac{1}{1+e^{\overline{y_{p_i}^{g,l}}}}\right)\right] \tag{7}$$

Furthermore, $\mathcal{L}_D$ aggregates the centrifugal potential loss functions of all decoders, as shown in (8).

$$\mathcal{L}_D = \frac{1}{(L-1)^2}\sum_{g=1}^{L-1}\sum_{l=2}^{L}\frac{1}{N^l}\sum_{i=1}^{N^l}\mathcal{L}_D^{g,l}(p_i) \tag{8}$$

Ultimately, the loss function for the entire network is given by (9).

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_D + \sum_{g=1}^{L-1}\mathcal{L}_{CE}^{g} \tag{9}$$

Where $\mathcal{L}_{CE}^{g}$ represents the original PointNeXt loss function, specifically the cross-entropy loss between the one-hot encoded labels from the final layer of each decoder and the ground-truth labels of the input point cloud. $\mathcal{L}$ regularizes the feature dispersion across multiple decoders and optimizes the discriminability of each decoder's feature map, which produces precise segmentation results.

3 Experiments

3.1 Experimental Implementation Details

We conducted training and testing of the network on the indoor point cloud dataset S3DIS, accepting point cloud inputs of 4096x6, which consists of 4096 points with 6 feature dimensions. The encoder progressively downsamples the number of centroid points to 160, while the feature dimension gradually increases to 512.

For the network training, considering the S3DIS dataset includes 6 indoor scenes and 11 categories, we applied 6-fold cross-validation to train the network. The batch size for network training is set to 8, while the batch size for the validation set is set to 1. In addition, we set the learning rate to 0.01 and apply cosine decay strategy to gradually reduce it to 1e-5.

3.2 Experimental results

We compared the performance of the PointNeXt-L network using contrastive learning with other point cloud semantic segmentation algorithms, as shown in Table 1. Our network achieved an mIoU of 75.3% and an OA of 90.2% on the S3DIS dataset using 6-fold cross-validation. On Area 5, the improved network incorporating contrastive learning achieved an mIoU of 71.03% and an OA of 90.3%.

Table 1. Experimental results on the S3DIS dataset.

Method	6-fold		Area5		Params (M)
	mIoU(%)	OA(%)	mIoU(%)	OA(%)	
PointNet [5]	47.7	78.6	43.7	78.9	1.7
PointNet++ [6]	54.5	81.0	53.5	83.0	-
SPG [7]	62.1	85.5	58.0	86.4	0.3
DeepGCN [8]	60.0	85.9	52.5	-	-
A-CNN [9]	62.9	87.3	-	87.3	-
MinkowskiNet [10]	65.4	-	65.4	-	-
PointCNN [11]	65.4	88.1	57.3	85.9	11.0
PointWeb [12]	66.7	87.3	60.3	87.0	-
PointASNL [13]	68.7	88.8	-	-	22.4
RandLA-Net [14]	70.0	87.1	62.5	-	1.2
PointSIFT [15]	70.2	88.7	-	-	-
KPConv [16]	70.6	-	67.1	-	14.9
SCF-Net [17]	71.6	88.4	63.4	-	-
BAAF-Net [18]	72.2	88.9	65.4	-	-
PointNeXt-L(baseline) [4]	73.9	89.9	69.0	90.0	7.1
PointNeXt-XL [4]	74.9	90.3	70.5	90.6	41.6
PointNeXt-L + $\mathcal{L}$ (ours)	75.3	90.2	71.0	91.8	13.3

3.3 Visualization results

We present the segmentation results of our network and the PointNext network on the S3DIS dataset, as shown in Figure 3. Our network outperforms PointNeXt in the segmentation of objects such as whiteboards and doors, demonstrating that contrastive learning improves the normalized representation of features for small-scale objects.

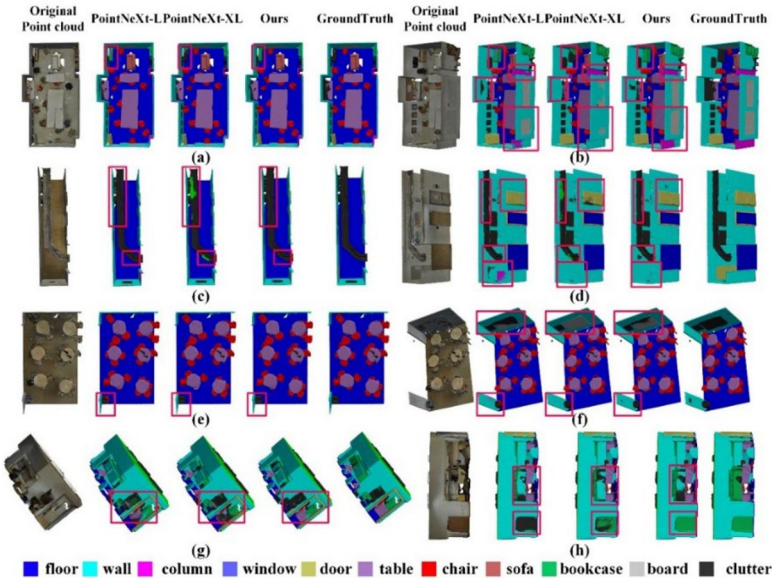

Fig. 3. Comparison of visualization results.

4 Conclusion

For indoor datasets, we propose a multi-scale supervised point cloud semantic segmentation network based on contrastive learning, which utilizes multi-hot labels generated by the encoder to supervise the multi-hot labels predicted by all decoders, and produces refined semantic segmentation results. Furthermore, we will fine-tune the indoor point cloud semantic segmentation model to enable it to accurately segment point cloud data from cockpits, thereby facilitating the generation of virtual cockpits [32].

References

1. Vlasblom, Jeanine, et al.: Virtual cockpit: making natural interaction possible in a low-cost VR simulator. In Proceedings of the 2021 IEEE Conference on Virtual Reality and 3D User Interfaces, vol. 6, pp. 1-12. IEEE Xplore (2021)
2. Sun, Y., Zhang, X., & Miao, Y.: A review of point cloud segmentation for understanding 3D indoor scenes. Visual Intelligence, vol. 2, pp. 14. Springer (2024) https://doi.org/10.1007/s44267-024-00046-x
3. Armeni I, Sener O et al.: 3D Semantic Parsing of Large-Scale Indoor Spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1534-1543. Las Vegas (2016) https://cvpr2016.thecvf.com
4. Qian, Guocheng, et al.: Pointnext: Revisiting pointnet++ with improved training and scaling strategies. Advances in neural information processing systems. vol.35. pp. 23192-23204. Curran Associates Inc. (2022) https://doi.org/10.48550/arXiv.2206.04670

5. Zhang, Jiaying, et al.: A review of deep learning-based semantic segmentation for point cloud. IEEE access **7,** 179118-179133. (2019) https://doi.org/10.1109/ACCESS.2019.2959756
6. Qi, Charles Ruizhongtai, et al.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems, vol. 30. (2017) https://doi.org/10.48550/arXiv.1706.02413
7. Landrieu, Loic, and Martin Simonovsky.: Large-scale point cloud semantic segmentation with superpoint graphs. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). (2018) https://doi.org/10.1109/CVPR.2018.00479
8. Guohao, Li, et al.: Deepgcns: Making gcns go as deep as cnns. IEEE transactions on pattern analysis and machine intelligence (T-PAMI). IEEE. (2021) https://doi.org/10.1109/TPAMI.2021.3074057
9. Komarichev, Artem, Zichun Zhong, and Jing Hua. A-cnn: Annularly convolutional neural networks on point clouds. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) https://doi.org/10.1109/CVPR.2019.00760
10. Choy, Christopher, 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 (CVPR). (2019) https://doi.org/10.1109/CVPR.2019.00319
11. Li, Yangyan, et al.: Pointcnn: Convolution on x-transformed points. Advances in neural information processing systems, vol.31. (2018) https://doi.org/10.48550/arXiv.1801.07791
12. Zhao, Hengshuang, et al.: Pointweb: Enhancing local neighborhood features for point cloud processing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) https://doi.org/10.1109/CVPR.2019.00571
13. Yan, Xu, et al.: Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) https://doi.org/10.1109/CVPR42600.2020.00563
14. Hu, Qingyong, et al.: Randla-net: Efficient semantic segmentation of large-scale point clouds. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) https://doi.org/10.1109/CVPR42600.2020.01112
15. Jiang, Mingyang, et al.: Pointsift: A sift-like network module for 3d point cloud semantic segmentation. arXiv preprint arXiv:1807.00652. (2018) https://doi.org/10.48550/arXiv.1807.00652
16. Fan, Siqi, et al.: SCF-Net: Learning spatial contextual features for large-scale point cloud segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2021) https://doi.org/10.1109/CVPR46437.2021.01427
17. Qiu, Shi, Saeed Anwar, and Nick Barnes.: Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2021) https://doi.org/10.1109/CVPR46437.2021.00180
18. Tang, L., Zhan, Y., Chen, Z., Yu, B., Tao, D.: Contrastive Boundary Learning for Point Cloud Segmentation. CVPR, pp.8479–8489. (2022) https://doi.org/10.1109/CVPR52688.2022.00830
19. Afham, M., Dissanayake, I., Dissanayake, D., Dharmasiri, A., Thilakarathna, K., Rodrigo, R.: CrossPoint: Self-Supervised Cross-Modal Contrastive Learning for 3D Point Cloud Understanding. CVPR, pp.9892–9902. (2022) https://doi.org/10.1109/CVPR52688.2022.00967

20. Yin, Y., Liu, Y., Xiao, Y., Cohen-Or, D., Huang, J., Chen, B.: SAI3D: Segment Any Instance in 3D Scenes. CVPR, pp.17144–17154. (2024) https://doi.org/10.1109/CVPR52733.2024.00317
21. Pissas, T., Ravasio, C. S., Da Cruz, L., Bergeles, C.: Multi-scale and Cross-scale Contrastive Learning for Semantic Segmentation. Lecture Notes in Computer Science, vol.13689, pp.413–429. (2022) https://doi.org/10.1007/978-3-031-19818-2_24
22. Xiao, A., Zhang, X., Shao, L., Lu, S.: A Survey of Label-Efficient Deep Learning for 3D Point Clouds. IEEE TPAMI, vol.46(12), pp.9139–9160. (2024) https://doi.org/10.1109/TPAMI.2024.3416302
23. Ye, S., Chen, D., Han, S., Liao, J.: Robust Point Cloud Segmentation With Noisy Annotations. IEEE TPAMI, vol.45(6), pp.7696–7710. (2023) https://doi.org/10.1109/TPAMI.2022.3225323
24. Zhang, W.: Real-Time Semantic Segmentation of Point Clouds Based on an Attention Mechanism and a Sparse Tensor. Applied Sciences, vol.13(5), pp.3256. (2023) https://doi.org/10.3390/app13053256
25. Yang, J., Chen, T.: Cross-Modal Contrastive Learning for Domain Adaptation in 3D Semantic Segmentation. AAAI, vol.37(3), pp.2974–2982. (2023) https://doi.org/10.1609/aaai.v37i3.25400
26. Won, J.S., Han, J.K.: Point Cloud Upsampling for Accurate Surface Reconstruction via Attention-Guided Generation. IEEE Access, vol.13, pp.172626–172637. (2025) https://doi.org/10.1109/ACCESS.2025.3616323
27. Humblot-Renaux, G., Jensen, S.B., Møgelmose, A.: From CAD Models to Soft Point Cloud Labels: An Automatic Annotation Pipeline for Cheaply Supervised 3D Semantic Segmentation. Remote Sensing, vol.15(14), pp.3578. (2023) https://doi.org/10.3390/rs15143578
28. Zhang, Y., Yao, J., Zhang, R., Chen, S., Fu, H.: HAVANA: Hard Negative Sample-Aware Self-Supervised Contrastive Learning for Airborne Laser Scanning Point Cloud Semantic Segmentation. Remote Sensing, vol.16(3), pp.485. (2024) https://doi.org/10.3390/rs16030485
29. Li, M., Li, M.F., Li, M., Xu, L.H.: Building Lightweight 3D Indoor Models from Point Clouds with Enhanced Scene Understanding. Remote Sensing, vol.17(4), pp.596. (2025) https://doi.org/10.3390/rs17040596
30. Li, S., Zheng, D., Zou, X.: Hierarchical Local–Global Context Modeling with Selective Modulation for Point Cloud Semantic Segmentation. ICMR, pp.1035–1044. (2026) https://doi.org/10.1145/3805622.3810831
31. Miao, Y.W., Sun, Y.L., Zhang, Y.M., Wang, J.R., Zhang, X.D.: An efficient point cloud semantic segmentation network with multiscale super-patch transformer. Scientific Reports, vol.14(1), pp.14581. (2024) https://doi.org/10.1038/s41598-024-63451-8
32. Kalscheuer, F., Eschen, H., Schüppstuhl, T.: Towards Semi-Automated Pre-assembly for Aircraft Interior Production. Advances in Assembly and Disassembly, pp.203–213. (2022) https://doi.org/10.1007/978-3-030-74032-0_17