Text-Guided Semantic Segmentation Method for Indoor 3D Point Clouds

Jinyu Tan[1], Juntao Yang[1,*], Yutao Zhang[1], Sa Li[1], Dandan Liu[1], Zhengwen Wang[1], Xue Zhang[2]

[1] College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China
tjy2424904994@163.com, jtyang@sdust.edu.cn, zhangyutao060116@outlook.com, 744853961@qq.com, 1870739192@qq.com, 15275571210@163.com
[2] College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China - xuezhang@sdust.edu.cn

Keywords: 3D point clouds, semantic segmentation, text prototypes, cross-modal learning, 3D scene understanding

Abstract: Point cloud semantic segmentation of indoor environments is a fundamental task in 3D scene understanding. However, existing methods mainly rely on geometric structures and color information, which are prone to error results in scenarios involving occlusion, sparse sampling, and geometrically similar structures. To address this issue, this paper proposes a text-knowledge-guided method for the point cloud semantic segmentation of indoor 3D scene. Built upon RandLA-Net as the baseline, the proposed method first constructs the textual semantic prototypes using multi-template prompts, and further enhances the stability of semantic anchors through periodic prototype refreshing. Then, a cross-modal semantic feature alignment mechanism is introduced at both the shallow and the high-level feature stages. Through feature alignment, bidirectional semantic interaction, and gated fusion, textual priors are progressively injected into the point cloud feature learning process. Finally, the model is jointly trained with a point-wise classification loss, a text-prototype alignment constraint, and a boundary optimization constraint to improve the semantic feature discrimination and segmentation boundary quality. Experimental results on the S3DIS dataset demonstrate that the proposed method achieves 86.8% OA, 81.6% mAcc, and 67.2% mIoU, exhibiting more stable segmentation performance in complex indoor scenes. As a consequence, these results indicate that incorporating textual semantic priors can effectively enhance high-level semantic representations of point clouds, providing a feasible solution for indoor 3D scene understand.

1. Introduction

Point cloud semantic segmentation of indoor environments aims to assign a semantic label to each point in a scene. As a fundamental task in 3D scene understanding, it plays an important role in applications such as robot navigation, digital twins, and building information modeling. Existing methods mainly rely on the geometric structure and color information of point clouds, which are prone to error results in scenarios involving occlusion, sparse sampling, and categories with similar geometric shapes (Charles R Qi et al., 2017; Charles Ruizhongtai Qi et al., 2017; Wang et al., 2019). To address this issue, this paper proposes a text-knowledge-guided method for indoor 3D scene point cloud semantic segmentation(Thomas et al., 2019; Zhao et al., 2021). Built upon RandLA-Net as the baseline, the proposed method constructs textual semantic prototypes through multi-template prompting, and introduces a cross-modal semantic enhancement mechanism at both the shallow and high-level feature stages(Qian et al., 2022). Combined with segmentation supervision, prototype alignment, and boundary optimization in a joint training framework, the proposed method improves the point-wise semantic discrimination capability in complex indoor scenes. The main contributions of this paper are summarized as follows:

(1) A text-knowledge-guided framework for indoor 3D scene point cloud semantic segmentation is proposed. By introducing category-level semantic priors into the RandLA-Net baseline, the proposed framework enhances the model’s ability to distinguish semantic categories in complex scenes.

(2) A textual semantic prototype construction method based on multi-template prompting is designed, where the prototype updating mechanism is introduced to improve the stability and robustness of textual semantic anchors.

(3) A cross-modal semantic feature alignment strategy is developed to integrate point cloud features with textual semantic features. Combined with segmentation supervision, semantic alignment constraints, and a boundary optimization objective, the proposed method further improves the overall performance of indoor point cloud semantic segmentation.

2. Experimental Method

2.1 Dataset and Evaluation Metrics

To validate the effectiveness and generalization capability of the proposed method, experiments were conducted on the public Stanford Large-Scale 3D Indoor Spaces (S3DIS) dataset(Armeni et al., 2016). This dataset contains six independent indoor areas, covering typical scenes such as offices, conference rooms, and corridors, with more than 215 million semantically annotated points in total. All point clouds are annotated into 13 semantic categories, and each point is associated with 3D coordinates and RGB color information, which jointly describe both geometric structure and appearance characteristics. To ensure fair comparison with existing studies, the official six-fold cross-validation protocol was adopted. In each fold, one area was used for testing, while the remaining five areas were used for training. This evaluation setting reduces the influence of distribution differences across individual areas on the experimental results.

To comprehensively evaluate segmentation performance, several quantitative metrics were employed, including class-wise Intersection over Union (IoU), mean Intersection over Union (mIoU), Overall Accuracy (OA), mean Accuracy (mAcc), and the Kappa coefficient. Their definitions are given in Eqs. (1) to

* Corresponding author: jtyang@sdust.edu.cn

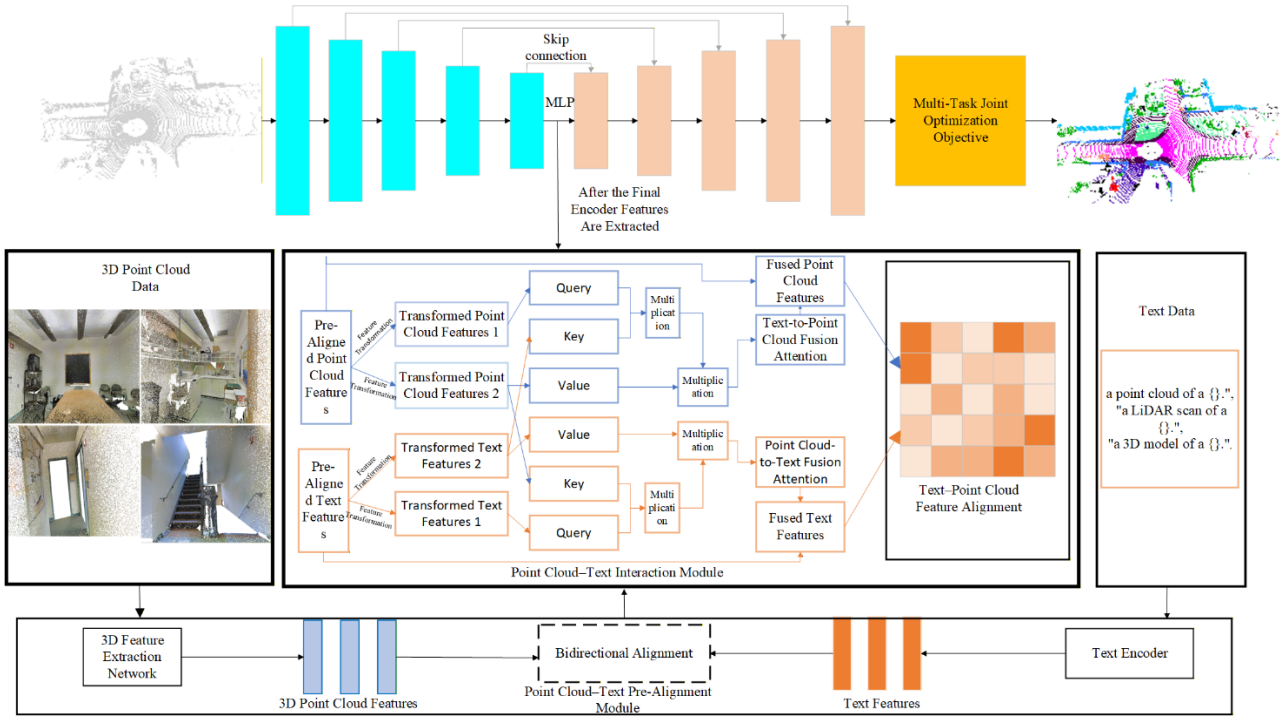

Figure 1. Overall pipeline of the proposed text-knowledge-guided network for indoor 3D scene point cloud semantic segmentation

(5). Specifically, IoU measures the overlap between the predicted region and the ground-truth region for each class, and can simultaneously penalize false positives and false negatives. mIoU is the average IoU over all categories and is used to reflect the overall performance of the model under class-imbalanced conditions. OA denotes the proportion of correctly predicted points among all points, while mAcc represents the average classification accuracy across categories. The Kappa coefficient is further adopted to evaluate the consistency between prediction results and ground-truth annotations.

$$\mathrm{IoU} = \frac{TP_i}{GT_i + FP_i} \tag{1}$$

$$\mathrm{mIoU} = \frac{\sum_{i=1}^{C} \mathrm{IoU}_i}{C} \tag{2}$$

$$Overall\ accuracy\ (OA) = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$mAcc = \frac{1}{C}\sum_{i=1}^{C} \frac{TP_i}{GT_i} \tag{4}$$

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

Where, TP denotes the number of positive samples correctly classified as positive, TN denotes the number of negative samples correctly classified as negative, FN denotes the number of positive samples incorrectly classified as negative, and FP denotes the number of negative samples incorrectly classified as positive. TP_i, GT_i, and FP_i represent the number of correctly predicted positive samples, the number of ground-truth samples, and the number of negative samples incorrectly predicted as positive for the i-th class, respectively. p_o denotes the overall accuracy, while p_e can be expressed as:

$$p_e = \frac{\sum_{i=1}^{C} a_{i+} * a_{+i}}{N^2} \tag{6}$$

Where a_{i+} and a_{+i} denote the numbers of ground-truth and predicted samples for the i-th class, respectively, and N denotes the total number of samples.

2.2 Overall Framework

The proposed text-knowledge-guided method for indoor 3D scene point cloud semantic segmentation adopts an encoder-decoder architecture, with RandLA-Net serving as the baseline (Hu et al., 2020). Given an input point cloud, each point is represented by its 3D coordinates and RGB color information. The input point cloud is first transformed by a feature mapping layer for initial feature embedding, and is then fed into multiple dilated residual encoding modules. During the progressive random downsampling process, the receptive field is continuously enlarged, enabling the extraction of hierarchical geometric-semantic features. After encoding, the network obtains high-level point cloud features with strong abstract semantic representations. Subsequently, the decoder gradually restores point-wise spatial resolution through layer-wise interpolation and skip connections, and outputs the category prediction for each point. Different from conventional point cloud segmentation methods that rely solely on geometric structure and color information, the proposed method introduces textual knowledge priors into the point cloud feature learning process to enhance the model’s discriminative capability (Ding et al., 2023; Peng et al., 2023). Specifically, two types of textual information are incorporated during training. The first is textual prototypes constructed from category lables, which provide stable semantic anchors for point-wise features. The second is auxiliary scene description text, which provides scene-level semantic context. The former is mainly used for point-wise semantic alignment supervision, while the latter is primarily employed for cross-modal semantic feature alignment.

Based on the above design, the overall pipeline of the proposed method can be summarized as follows. First, textual semantic prototypes and scene description text are constructed. Next, a text encoder is used to map the textual inputs into the semantic feature space. The textual semantic information is then progressively injected into the point cloud encoding stage and the high-level feature fusion stage. Finally, a joint optimization objective is adopted to collaboratively constrain point-wise classification results, category-level semantic consistency, and boundary quality. The overall framework is illustrated in Figure 1.

Table 1. Comparison of different methods on the S3DIS dataset

Methods	OA	mAcc(%)	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book	board	clut.
PointNet	78.6	66.2	47.6	88.0	88.7	69.3	42.4	23.1	47.5	51.6	54.1	42.0	9.6	38.2	29.4	35.2
RSNet	-	66.5	56.5	92.5	92.8	78.6	32.8	34.4	51.6	68.1	59.7	60.1	16.4	50.2	44.9	52.0
SPG	85.5	73.0	62.1	89.9	95.1	76.4	62.8	47.1	55.3	68.4	73.5	69.2	63.2	45.9	8.7	52.9
PointCNN	88.1	75.6	65.4	94.8	97.3	75.8	63.3	51.7	58.4	57.2	71.6	69.1	39.1	61.2	52.2	58.6
PointWeb	87.3	76.2	66.7	93.5	94.2	80.8	52.4	41.3	64.9	68.1	71.4	67.1	50.3	62.7	62.2	58.5
ShellNet	87.1	-	66.8	90.2	93.6	79.9	60.4	44.1	64.9	52.9	71.6	84.7	53.8	64.6	48.6	59.4
Ours	86.8	81.6	67.2	92.4	95.4	78.5	50.5	53.6	61.4	65.9	65.3	77.5	54.8	61.4	61.3	56.3

2.3 Construction of Textual Semantic Information

2.3.1 Construction of Textual Semantic Prototypes: To enhance the stability of textual priors, a multi-template prompting strategy is adopted to construct textual semantic prototypes. For each category, instead of using only a single category label as the textual input, multiple descriptive templates are designed, and the category label is inserted into different sentence patterns to generate a set of semantic prompts (Radford et al., 2021; Zhou et al., 2022a; Zhou et al., 2022b). In this way, the semantic bias caused by a single expression form can be reduced, making the textual representation more robust(Devlin et al., 2019). Let $t_{(c,m)}$denote the textual feature of the c-th category under the m-th template. After being fed into a pre-trained text encoder, the corresponding token-level textual features are obtained. Considering that different text sequences may have different lengths, masked average pooling is adopted to obtain the sentence-level semantic representation, and the textual feature generated from each template is L2-normalized. Subsequently, the textual features of all templates belonging to the same category are averaged, and the resulting feature is normalized again to obtain the final semantic prototype, which can be expressed as follows:

$$p_c = \mathrm{Norm}(\frac{1}{M}\sum_{m=1}^{M}\mathrm{Norm}(\mathrm{Pool}\left(E\left(t_{c,m}\right)\right))) \quad (7)$$

Where, Norm $(\cdot)$ denotes L2 normalization, M is the number of templates, $E(\cdot)$ represents the text encoder, and Pool $(\cdot)$ denotes the masked average pooling operation.

2.3.2 Construction of Scene Description Text: In addition to textual prototypes, auxiliary scene description text is also constructed to provide scene-level semantic context for point cloud feature learning. Specifically, under the current experimental setting, a short textual description is organized according to the annotated categories corresponding to each input point cloud, and is then fed into the text encoder to obtain token-level textual features. This textual feature mainly serves as a label-assisted cross-modal semantic guidance signal to enhance the alignment between point cloud features and semantic information.

2.4 Text-Knowledge-Guided Point Cloud Segmentation Network

2.4.1 Point Cloud Feature Encoding and Shallow Semantic Modulation: Given an input point cloud, the network first employs the RandLA-Net to extract hierarchical point cloud features. At the shallow feature extraction stage, local geometric structures and color information of the point cloud still dominate the feature representation. However, appropriately injecting textual semantic priors at this stage can improve the network's sensitivity to category-related information. Based on this consideration, the global vector of the scene description is used to conditionally modulate the output features of the first encoder layer.

Let z denote the global vector obtained from the encoded scene description, and let F_1 denote the point cloud feature output by the first encoder layer. Two lightweight mapping branches are then employed to generate the scaling term and bias term associated with z, respectively, for feature modulation of F_1, which can be formulated as follows:

$$F_1' = F_1 \odot \left(1 + \alpha S(z)\right) + \alpha B(z) \quad (8)$$

Where, $S(\cdot)$ and $B(\cdot)$ denote the scaling mapping branch and the bias mapping branch, respectively, α is a learnable coefficient, and $\odot$ denotes element-wise multiplication. Through this mechanism, the network is able to introduce sceney-related semantic bias while preserving shallow geometric detail information, thereby providing more explicit semantic guidance for subsequent deep feature learning.

2.4.2 High-Level Cross-Modal Semantic Interaction and Fusion: After obtaining the bottleneck features, a cross-modal semantic feature alignment module is introduced in the high-level semantic space. Let $P \in \mathbb{R}^{B\times N\times D}$ denote the high-level point cloud features and $T \in \mathbb{R}^{B\times L\times D}$ denote the token-level text features, where B, N, L, and D are the batch size, number of points, number of text tokens, and feature dimension, respectively. To reduce the modality gap, a lightweight bidirectional alignment module is first applied, followed by bidirectional cross-modal attention to model point-to-text and text-to-point interactions. This enables point cloud features to absorb semantically relevant textual cues while refining text features through point cloud responses.

To avoid introducing textual noise and weakening geometric structure information, an adaptive gated fusion mechanism is used to combine the original point cloud features with the interacted features P_f:

$$G = \sigma\left(\mathrm{MLP}\left(\left[P_f; P\right]\right)\right) \quad (9)$$

$$P' = G \odot P_f + (1 - G) \odot P \quad (10)$$

Where $[\,\cdot\,;\,\cdot\,]$denotes feature concatenation, MLP is a multi-layer perceptron, σ is the Sigmoid function, and G is the gating weight. In this way, textual semantics are adaptively injected while preserving geometric structure. The fused features are then sent to the decoder to produce the final point-wise segmentation results.

2.5 Joint Optimization Objective

To simultaneously ensure point-wise classification accuracy, category-level semantic consistency, and segmentation boundary quality, a multi-loss joint optimization strategy is adopted for training. First, the cross-entropy loss is used to supervise the classification results of each valid point, so as to guarantee the basic point-wise semantic segmentation capability of the model, which is denoted as L_{seg}. To further improve the IoU metric and enhance prediction quality in category boundary regions(Berman et al., 2018), the Lovasz-Softmax loss is introduced and denoted as L_{lovasz}. This loss provides an additional constraint on

Table 2. Six-fold cross-validation results for semantic segmentation on the S3DIS dataset

Testing area	OA	mAcc(%)	mIoU(%)	Kappa	ceil.	Floor	wall	beam	col.	wind.	door	table	chair	sofa	book-case	board	clut.
Area 1	87.5	85.2	72.4	0.853	96.3	94.6	75.3	67.2	51.2	82.6	78.8	67.8	80.7	63.1	56.0	64.4	62.9
Area 2	83.8	76.3	56.6	0.805	86.5	94.5	78.2	39.8	63.3	52.7	61.0	40.0	71.8	23.4	46.6	36.6	41.3
Area 3	89.6	88.7	76.1	0.876	93.8	96.0	78.6	72.2	41.7	75.1	79.9	70.9	77.6	75.4	76.2	82.9	68.8
Area 4	84.5	78.8	61.3	0.814	93.6	95.7	77.9	42.7	62.5	30.2	61.5	63.6	69.6	44.5	51.8	51.4	49.7
Area 5	84.4	69.4	58.8	0.811	87.9	93.9	79.2	0.0	34.8	57.0	28.5	74.4	81.7	60.2	63.4	57.7	46.1
Area 6	90.9	91.4	78.1	0.894	96.4	97.5	81.9	81.6	67.9	70.8	85.9	74.9	83.8	61.9	74.5	74.5	68.7

the classification results from the perspective of IoU optimization, and is particularly beneficial for improving segmentation quality at boundary points and in hard-to-distinguish regions.

To ensure that point-wise features remain semantically consistent with the corresponding textual prototypes in the shared semantic space, a semantic prototype alignment loss is further introduced. Let the projected point-wise feature be denoted by f_i, its corresponding ground-truth category by y_i, and the textual prototype of the c-th category by p_c. Then, the semantic matching degree between the point-wise feature and the textual prototypes of all categories is computed using normalized cosine similarity as follows:

$$s_{(i,c)} = \frac{\mathrm{Norm}(f_i)^T \mathrm{Norm}(p_c)}{\tau} \tag{11}$$

Where, τ denotes the temperature coefficient. Subsequently, the similarity scores are treated as classification logits and combined with the ground-truth labels to compute the cross-entropy loss, yielding the semantic prototype alignment loss L_{align}. This constraint encourages point features belonging to the same category to cluster around the corresponding textual prototype, while enlarging the semantic margin between different categories. Considering that some semantic categories in indoor scenes exhibit high similarity in geometric structure and appearance, the model may produce persistent confusion among these categories. To further enhance the discriminability between easily confused categories, a pairwise confusion-margin loss L_{pair} is introduced. For predefined pairs of confusing categories, this loss enforces the score of the ground-truth category to be higher than that of the confusing category by a certain margin, thereby reducing misclassification between similar categories. By integrating the above loss terms, the overall loss function of this paper can be formulated as follows:

$$L = L_{\text{seg}} + \lambda_1 L_{\text{align}} + \lambda_2 L_{\text{lovasz}} + \lambda_3 L_{\text{pair}} \tag{12}$$

Where, λ_1, λ_2, and λ_3 denote the weighting coefficients of the category prototype alignment loss, the Lovasz loss, and the pairwise confusion-margin loss, respectively. Through the above joint optimization strategy, the model is able to learn point-wise feature representations that simultaneously account for local geometric structure, high-level semantic priors, and boundary discrimination capability, thereby improving the overall performance of indoor 3D scene point cloud semantic segmentation.

3. Experimental Analysis

3.1 Experimental Details

The experiments were implemented using the PyTorch deep learning framework on an Ubuntu 20.04 system. Model training was conducted on a 32 GB vGPU, and the network was trained for 100 epochs in total. During training, the batch size was set to 4 by default. The Adam optimizer was adopted, with a weight decay of 1×10^{-4}. The initial learning rate was set to 1×10^{-2}, and an exponential learning rate scheduler with $\gamma = 0.95$was employed for decay to maintain stable optimization in the later training stage. In addition, to alleviate overfitting, a Dropout layer with a dropout rate of 0.5 was inserted before the classification head.

3.2 Quantitative Comparison on the S3DIS Dataset

To verify the effectiveness of the proposed method for indoor 3D scene point cloud semantic segmentation, the proposed approach was compared with several representative point cloud segmentation models on the S3DIS dataset. The comparison results are presented in Table 1. The baseline methods include representative models such as PointNet, RSNet, SPG, PointCNN, PointWeb, and ShellNet (Zhang et al., 2019; Zhao et al., 2019). The evaluation metrics include OA, mAcc, mIoU, and the IoU for each category (Huang et al., 2018; Landrieu and Simonovsky, 2018; Li et al., 2018).

Table 1 compares the proposed method with several classical point cloud segmentation models on the S3DIS dataset. The proposed method achieves 86.8% OA, 81.6% mAcc, and 67.2% mIoU. Both mAcc and mIoU are higher than those of PointCNN, PointWeb, and ShellNet, indicating better class balance and overall segmentation performance. In terms of class-wise IoU, the proposed method performs well on major structural categories such as ceiling, floor, and wall, and also achieves favorable results on challenging categories including column, chair, bookcase, and board. These results suggest that text-knowledge guidance improves the model’s discriminative ability for complex categories.

3.3 Analysis of Six-Fold Cross-Validation Results

Table 2 presents the detailed results for the six test areas. It can be observed that the model performance varies across different areas to some extent, while maintaining relatively good overall stability. Among them, Area 6 achieves the best performance, with an mIoU of 78.1%; Area 3 and Area 1 also show favorable results. In contrast, the segmentation performance on Area 2, Area 4, and Area 5 is relatively lower, indicating that there is still room for improvement in scenarios with larger cross-area distribution differences and a greater number of weak-category targets.

3.4 Confusion Matrix Analysis

The confusion matrix and visualization results (Figure 2) show that the proposed method achieves high recognition accuracy for large-area structural categories such as ceiling, floor, and wall, and also produces relatively stable predictions for typical furniture categories such as chair and table. However, some misclassification still exists among categories such as beam, column, window, as well as bookcase, board, and clutter. Overall, the text-knowledge-guided mechanism enhances the high-level semantic representation capability of point cloud features and alleviates the confusion among similar categories to a certain extent.

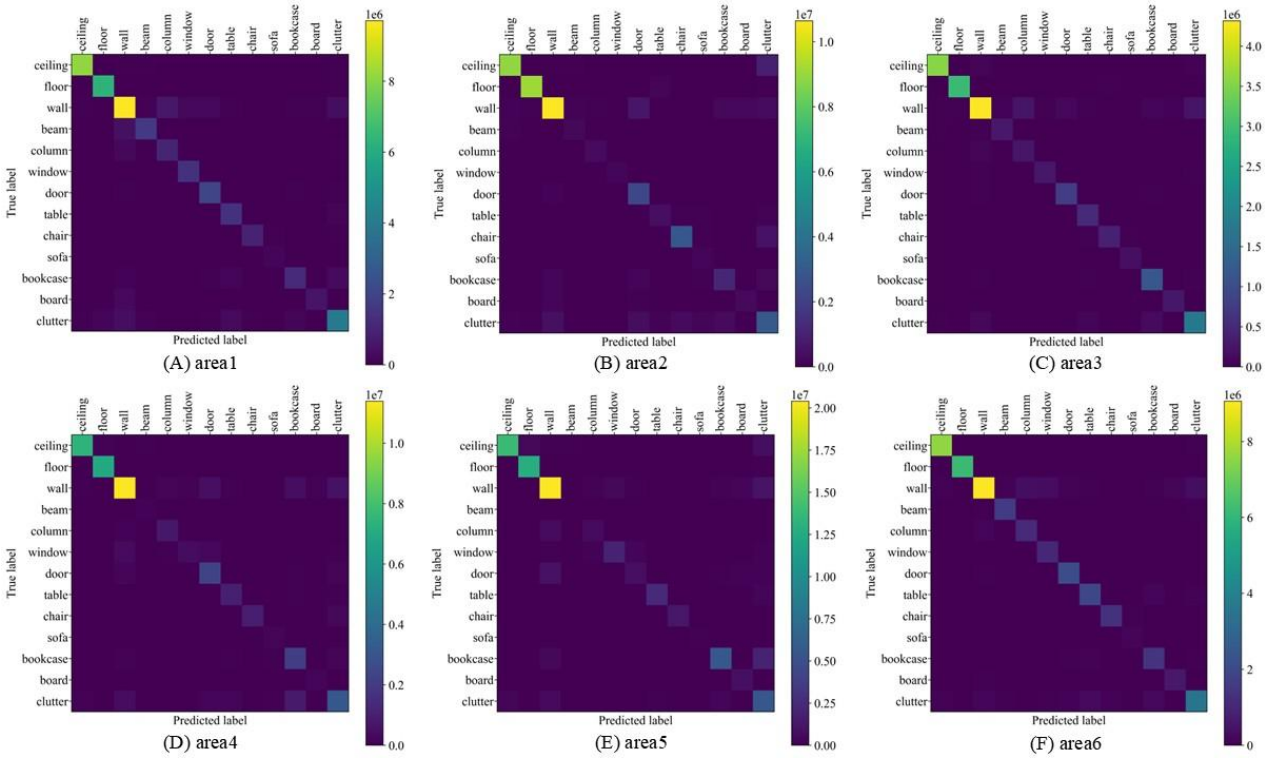

Figure 2. Visualization of the confusion matrices for the six areas of the S3DIS dataset

4. Conclusion

To address the limitation of insufficient high-level semantic modeling in indoor 3D scene point cloud semantic segmentation, this paper proposes a text-knowledge-guided semantic segmentation framework. The proposed method is built upon RandLA-Net, which serves as the baseline for efficient hierarchical point cloud feature extraction. On this basis, category-level semantic priors derived from a pre-trained text encoder are introduced to strengthen the semantic representation capability of the network. Specifically, textual semantic prototypes are constructed through multi-template prompting to improve the robustness and stability of textual category descriptions. In addition, a shallow semantic modulation strategy is employed to inject scene-related semantic cues into the early stage of feature learning, while a high-level cross-modal fusion mechanism is designed to enhance the interaction between point cloud features and textual semantic information in the deeper semantic space. To further improve segmentation performance, the model is trained under a joint optimization framework that integrates segmentation supervision, prototype alignment, and boundary refinement, enabling it to learn point-wise representations that simultaneously preserve local geometric details and encode high-level semantic knowledge.

Extensive experiments on the S3DIS dataset demonstrate the effectiveness of the proposed method. The model achieves 86.8% OA, 81.6% mAcc, and 67.2% mIoU, showing competitive overall performance for indoor scene segmentation. The results also indicate that introducing textual semantic priors can effectively enhance the discriminative ability of the network, particularly in complex indoor environments and for semantically similar categories that are difficult to separate using geometric and color cues alone. These findings suggest that text-guided semantic enhancement provides a feasible and promising direction for improving indoor 3D scene understanding. In future work, more stable strategies for textual prior construction and stronger cross-scene generalization capability will be further investigated, with the goal of improving the robustness and applicability of the proposed method in more diverse real-world scenarios.

Acknowledgements

This work is jointly supported by the National Natural Science Foundation of China (No. 42201486, 42371453, 62302278), the Natural Science Foundation of Shandong Province under Grant ZR2023QF014, the Qingdao Natural Science Foundation under Grant 23-2-1-112-zyyd-jch, the Higher Education Institutions Youth Innovation and Science & Technology Support Program of Shandong Province under Grant 2024KJH066, the Young Talent of Lifting engineering for Science and Technology in Shandong, China [SDAST2024QTA055].

References

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., & Savarese, S., 2016. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1534-1543)..

Berman, M., Triki, A. R., & Blaschko, M. B., 2018. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4413-4421).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., & Qi, X., 2023. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision*

and pattern recognition (pp. 7010-7019).

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., & Markham, A., 2020. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11108-11117).

Huang, Q., Wang, W., & Neumann, U., 2018. Recurrent slice networks for 3d segmentation of point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2626-2635).

Landrieu, L., & Simonovsky, M., 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4558-4567).

Li, Y., Bu, R., Sun, M., Wu, W., Di, X., & Chen, B. J. A. i. n. i. p. s., 2018. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, *31*.

Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., & Funkhouser, T., 2023. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 815-824).

Qi, C. R., Su, H., Mo, K., & Guibas, L. J., 2017. 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).

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. J. A. i. n. i. p. s., 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, *30*.

Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H., Elhoseiny, M., & Ghanem, B. J. A. i. n. i. p. s., 2022. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in neural information processing systems*, *35*, 23192-23204.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J., 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., & Guibas, L. J., 2019. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6411-6420).

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. J. A. T. o. G., 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, *38*(5), 1-12.

Zhang, Z., Hua, B.-S., & Yeung, S.-K., 2019. Shellnet: Efficient point cloud convolutional neural networks using concentric shells statistics. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1607-1616).

Zhao, H., Jiang, L., Fu, C.-W., & Jia, J., 2019. Pointweb: Enhancing local neighborhood features for point cloud processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5565-5573).

Zhao, H., Jiang, L., Jia, J., Torr, P. H., & Koltun, V., 2021. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 16259-16268).

Zhou, K., Yang, J., Loy, C. C., & Liu, Z., 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16816-16825).

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. J. I. j. o. c. v., 2022b. Learning to prompt for vision-language models. *International journal of computer vision*, *130*(9), 2337-2348.