Journal of Intelligent Manufacturing and Special Equipment

143

DSF-Net: semantic segmentation of large-scale point clouds based on integrating deep and shallow networks

Gang Xiao
College of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou, China
Yangsheng Zhong, Zhipeng Wang and Sihan Ge
College of Information Engineering, China Jiliang University, Hangzhou, China
Qibing Wang
College of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou, China
Feng Xu
Zhejiang Academy of Special Equipment Science, Hangzhou, China, and
Jiawei Lu
China Jiliang University, Hangzhou, China

Received 10 February 2025
Revised 23 March 2025
Accepted 11 April 2025

Abstract

Purpose – With the upgrading of three-dimensional (3D) sensing devices, the amount of point cloud data collected has also increased exponentially. However, most of the existing methods also have unbalanced optimizations in memory consumption and semantic segmentation efficiency. This research addresses the need for a more balanced approach in processing large-scale point cloud data efficiently.

Design/methodology/approach – This research used a network framework (DSF-Net) based on dual-path deep and shallow networks and designed a point cloud space pyramid pooling module based on hole convolution. The 3D point cloud data are trained separately by integrating the deep branch and shallow branch networks. Besides, a deep and shallow fusion module fuses the deep and shallow feature relationships and outputs several loss functions for convergence training.

Findings – It is found that DSF-Net solves the problem of segmentation efficiency, achieves a balanced effect while ensuring the ability of a large range of point cloud input and reduces the memory consumption.

Originality/value – The deep network can extract high-level semantic information, while the shallow neural network has fewer neural network layers and faster inference speed. Meanwhile, random sampling and point-atrous spatial pyramid pool modules are used, respectively, for deep and shallow networks to capture multi-scale local context information in point cloud.

Keywords Point cloud, Semantic segmentation, Atrous convolution, Pyramid pooling, Feature fusion

Paper type Research paper

1. Introduction

Point cloud is the predominant form of three-dimensional (3D) data and can be used for semantic segmentation. Semantic segmentation techniques applied to two-dimensional (2D)

This work is supported by the “Pioneer” and “Leading Goose” R&D Program of Zhejiang Province, China (Nos: 2025C01022 and 2023C01022), the LingYan Planning Project of Zhejiang Province, China (No: 2023C01215), and the Science and Technology Key Research Planning Project of HuZhou city, China (No: 2022ZD2019).

Journal of Intelligent Manufacturing and Special Equipment
Vol. 6 No. 2, 2025
pp. 143-154
Emerald Publishing Limited
e-ISSN: 2633-660X
p-ISSN: 2633-6596
DOI 10.1108/JIMSE-02-2025-0002

images are relatively mature, but there are still some challenges and bottlenecks in the semantic segmentation of 3D point clouds. In recent years, the semantic segmentation of 3D point clouds has been deeply studied, and the number of data sets used for indoor scanning by radar equipment has gradually increased (Gao *et al*., 2022; Jhaldiyal and Chaudhary, 2023; Shuai *et al*., 2021). However, once the amount of data increases, the efficiency of semantic segmentation decreases significantly. Besides, the lack of present 3D point cloud datasets for large scenes leads to limitations in the study of semantic segmentation of point clouds in large scenes. Large-scale point clouds refer to datasets composed of millions or even billions of 3D-coordinated points that can accurately express the geometric shapes, textures and environmental features of objects in space. They are typically derived from laser scanning, photogrammetry or sensor data. As the scale of the scene increases, the volume of point cloud data grows explosively, and the diversity and complexity of the scene increase the challenges associated with point cloud data. When processing point cloud data, issues such as noise, holes, occlusions and data loss are encountered.

Nowadays, work on semantic segmentation of point clouds is beginning to deal directly with large-scale point clouds, and these methods follow the whole-to-local principle of reducing the impact of the amount of data on the efficiency of segmentation from the point cloud sampling point of view. Recently, RandLA-Net used random sampling (Hu *et al*., 2020), which reduced memory usage and improved segmentation efficiency, but the original point cloud information may be lost, and these networks are trained on localized point clouds and could not be directly applied to large-field point clouds (Chen *et al*., 2022; Croitoru *et al*., 2023). Besides, DLA-Net proposed a neighborhood search algorithm and a local feature extraction module and completed the downsampling of the point cloud with a random sampling method (Su *et al*., 2022); multi-scale attentive aggregation (MSAA) proposed to assign the attention weights of different learning channels on the basis of local features, improved the splicing of contextual information and spliced the weights of high-level features with low-level features after compression (Geng *et al*., 2021). Although these methods reduce the memory usage and improve the computational efficiency, they have the problem of losing the structure of the original point cloud.

To address the aforementioned issues, a new framework is proposed that enables end-to-end training and segmentation of point clouds. First, the point clouds are first input as separate data sources into both the deep and shallow network architectures. The deep network receives point cloud data from random sampling, extracting contextual information, and the use of random sampling in the deep structure helps reduce memory consumption. In the same time, the shallow network receives point cloud data from the feature pyramid pooling layer, responsible for extracting spatial information, and utilizes a lightweight structure as a training network to balance the number of parameters. Finally, a fusion module is employed to combine the global and local refined features of the point clouds. In summary, our main contributions are as follows:

(1) A new framework is proposed that integrates deep learning of deep networks and shallow networks for semantic segmentation of large field data. Random sampling is used to reduce the memory usage during training. The deep network can extract high-level semantic information, while the shallow neural network has fewer neural network layers and a faster inference speed.

(2) According to the different responsibilities of deep and shallow networks, a random sampling and point-atrous spatial pyramid pool module is designed, respectively, for deep and shallow networks to capture multi-scale local context information in point cloud.

(3) Based on the above two aspects, a feature fusion module is introduced to construct a dual-path deep and shallow network for point cloud semantic segmentation. The experimental results on the S3DIS dataset and the SensatUrban dataset show that deep and shallow feature (DSF)-net has excellent performance. In addition, the effectiveness of DSF-net modules is validated by ablation experiments.

2. Relate work

2.1 Voxel-based segmentation

The super-voxels method first performs super-voxel segmentation of the point cloud and then performs feature extraction and classification for each super-voxels (Li *et al*., 2018). However, the results of super-voxel segmentation depend on the choice of hyperparameters, so hyperparameter tuning is required to achieve the best performance in different datasets and applications. The sparse voxel-based attention (SVA) method converts the point cloud data into a sparse set of voxels and uses a self-attention mechanism to capture the relationship between each voxel to improve the accuracy of semantic segmentation (Zhao *et al*., 2022). By inputting data sampled at different voxel resolutions simultaneously into the network, PVCFormer improves the segmentation of small-scale features while broadening the receptive field (Zhang *et al*., 2024). However, the information may be lost, which will affect the final semantic segmentation prediction. In general, point clouds lose too much information after voxelization.

2.2 Multi-view images-based segmentation

To convert 3D point cloud to 2D image for semantic segmentation, existing 2D segmentation algorithms can be used to reduce algorithm complexity (Lyu *et al*., 2020). The multi-view-based point cloud converts the point cloud into a set of fixed number of images and uses convolutional neural network to process these images, which can better capture the geometric structure and semantic information in the point cloud (Hamdi *et al*., 2021). The geometry-based multi-projection fusion module achieves the geometric feature alignment between range-view (RV) and bird-eye-view (BEV) and fuses the features of the two views at both feature level and output level (Xu *et al*., 2023). However, processing point cloud data may result in lost or inaccurate information. In summary, some geometric information may be lost when 3D point cloud data are converted to 2D image, so the accuracy of the algorithm may be affected.

2.3 Raw point cloud-based segmentation

The method of permutation invariance uses a max pooling operation in the feature extractor to summarize local features of each point into global features to deal with the disorder in point cloud (Liu *et al*., 2019, Qi *et al*., 2017a, b). Point-Bert can effectively learn semantic information and local structure in point cloud data by modeling mask points of point cloud data (Yu *et al*., 2022). LACV-Net seamlessly fuses local features from multiple resolution representations to capture global contextual features, generating a global descriptor vector (Zeng *et al*., 2024). However, it requires a lot of pre-training data and computational resources. In summary, point cloud data have many advantages and disadvantages, which need to be processed and analyzed with suitable algorithms and methods to achieve more accurate and reliable segmentation.

3. Method

In order to further balance the segmentation accuracy and efficiency of semantic segmentation under large-field attraction clouds, a bilateral architecture is designed, integrating deep and shallow networks, and used deep branch layers (D) and shallow branch layers (S), respectively, for semantic segmentation of point clouds, as shown in Figure 1. Firstly, the input irregular point cloud data is upgraded dimensional through the multi-Layer perceptron and used as the original input of the deep branch layer and the shallow branch layer, respectively.

Among them, the deep branch layer is composed of a relatively deep neural network, which can extract high-level semantic information. It selects a random sampling method to downsample the raw point cloud, which reduces the memory usage in the training process of the neural network. Since deep branch layer neural network has been trained for shallow branch layer, while ensuring the training efficiency, the module with small memory consumption and few training parameters for point cloud feature sampling. The point cloud

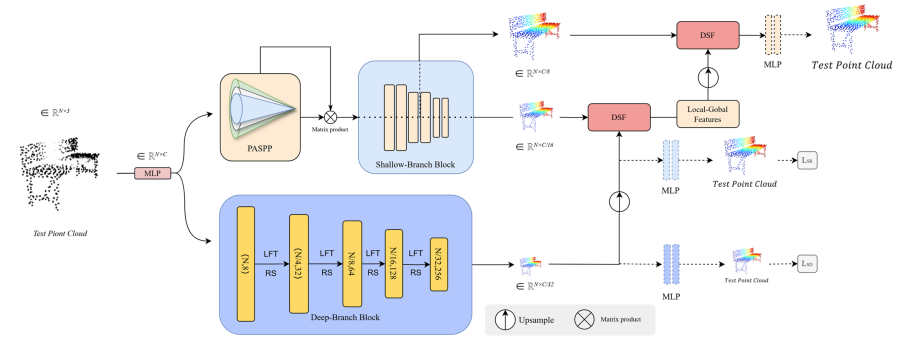

Figure 1. DSF-Net network model. (Source: Authors' own creation)

data processed by ascending dimension are input into the point-atrous spatial pyramid pooling (PASPP) module, and the processed point cloud is input into the shallow branch layer. Since the shallow branch layer has fewer neural network layers and faster inference speed, the input of this layer does not need downsampling operation. In order to better fuse this branch's features with the detailed shallow branch, an auxiliary segmentation head, a super-resolution head and a structure distillation loss are introduced to provide deep supervision. Three features are extracted from these two branches for feature fusion through DSF module.

3.1 Point-atrous spatial pyramid pooling (PASPP)

In the relevant research of ASPP, the problem of input sharing of different scale features can be solved by adding an ASPP layer after the convolution layer (Guo *et al.*, 2003). This method not only improves the feature sampling efficiency of the point cloud but also increases the receptive field of the network with less operational complexity. In order to improve the local feature loss and poor model generalization ability caused by MaxPooling operation in Pointnet++ (Charles Ruizhongtai Qi *et al.*, 2017a, b), PASPP is proposed applied in 3D space based on ASPP structure in 2D space.

Different from the common pooling convolution module, ASPP introduces the parameter $rate = x$ to represent the hyperparameter of the parameter interval in the convolution kernel. For pooling convolution kernel of different scales, the corresponding parameter intervals are also different. The PASPP module further uses multiple parallel 3D sampling layers with different sampling rates to explicitly capture multi-scale local context information in the point cloud and further process the features extracted at different sampling rates, as shown in Figure 2. The PASPP can be represented as follows:

$$F(f_1, f_2, \ldots f_n) = [p(f_1, k(s_1, x_1), \ldots, p(f_n, k(s_n, x_n)))] \tag{1}$$

Where f_n represents the high-dimensional feature information constructed through the MLP layer, p represents the MaxPooling operation, s_n represents the pooling window size and x_n represents the empty convolution step rate of the corresponding pooling window. Output F is the feature information extracted by the pyramid module.

3.2 Deep and shallow relation-aware feature fusion

3.2.1 Local location coding module. To fuse local feature information extracted from deep and shallow networks, k-nearest neighbor algorithm is used to generate a local neighborhood for location coding (Hu *et al.*, 2020). The specific process is as follows: for a given input with N point clouds, the 3D position information can be recorded as

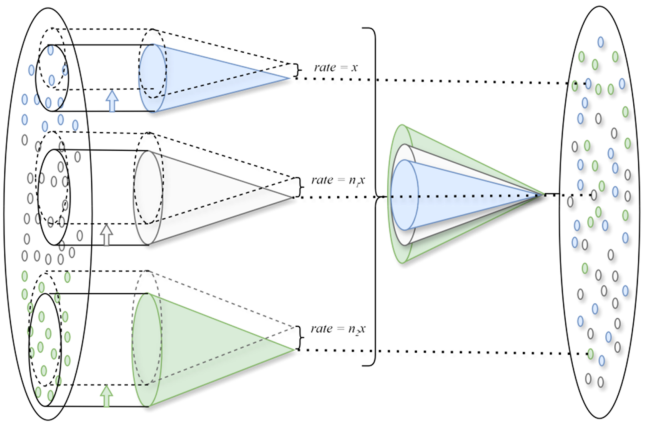

Figure 2. Point-atrous spatial pyramid pooling. (Source: Authors' own creation)

$N = \{N_1 \ldots N_i\} \in \mathbb{R}^3$ where any point N_i satisfies the neighborhood $\{N_i^1 \ldots N_i^K\} \in N$, the positional encoding can be represented as follows:

$$C_i^k = \left\{ \left(N_i - N_i^k\right) \oplus \left\| N_i - N_i^k \right\| \right\} \tag{2}$$

Where $\oplus$ is a join operation, and the $\|\cdot\|$ calculates the Euclidean distance between the adjacent point and the center point. Because the deep and shallow layer network is used to train the point cloud directly, other information such as red, green and blue (RGB) is filtered in the network. Therefore, in order to add positional coding features, 3D coordinates are used for *k*-nearest neighborhood search, and inputs positional features into the fusion model as parameters.

3.2.2 Relation-aware feature fusion. Aiming at the fusion of different feature dimensions of high-level semantic information in deep branches and detailed spatial information in shallow branches, the corresponding deep and shallow relation-aware feature fusion module (RAF) is introduced, as shown in Figure 3. Besides, the subtracter is selected as the relation function, and the location coding is added to the original feature and the relation function feature, respectively. The feature of deep and shallow layer network is set as $fi^k, f_i^k s \in \mathbb{R}^d, f_i^k{}_d \in \mathbb{R}^d$ (d is the feature channel), which represents the feature from shallow layer and deep layer network, respectively. The specific steps for feature fusion of deep and shallow layer network are as follows:

(1) Calculate the channeled attention parameter G:

$$G = \eta\left(\alpha(f_i) - \beta\left(f_i^k\right) + c_i^k\right) \tag{3}$$

Where f_i represents input point cloud information, α *and* β a $G =$ re MLP with a linear layer. The mapping function η is an MLP that contains two linear layers with ReLU activation. In addition, the two linear layers convert the feature to a higher dimension and back to the original dimension. Secondly, the relation matrix $\boldsymbol{R} \in \mathbb{R}_{d \times d}$ between G_s and G_d is defined by the inner product of each group:

$$R = G_s G_d^T \tag{4}$$

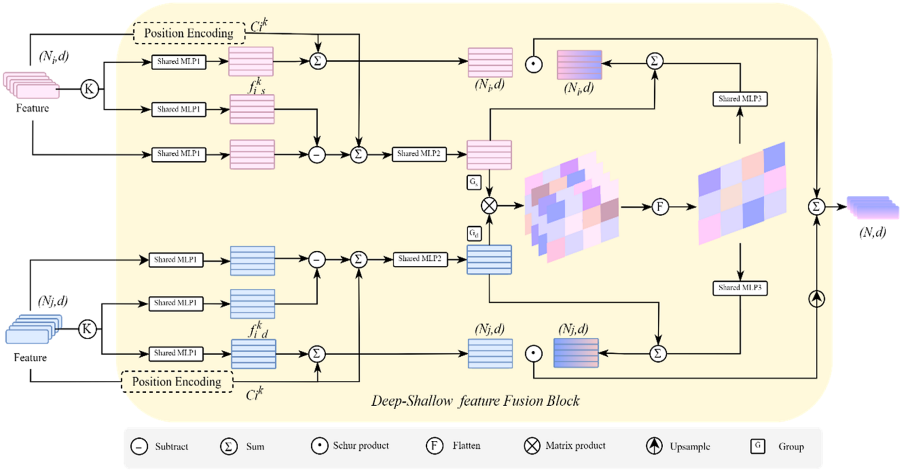

Figure 3. Deep and shallow relation-aware feature fusion (RAF). (Source: Authors' own creation)

(2) Calculate the modulation coefficient M:

$$M = \sigma(G + \mu fc(Flatten(R))) \tag{5}$$

Where σ is the sigmoid function and μ is the MLP with a linear layer.

(3) Calculate the characteristic factor coefficient Q:

$$Q = \gamma\left(f_i^k\right) + c_i^k \tag{6}$$

The feature factor is calculated as the sum of the parameters of the location feature and the point cloud feature, and the parameter can be used as the dot multiplier factor of the element vector of the fusion feature.

(4) Calculate RAF parameter F_{fusion}:

$$F_{fusion} = M_S \cdot Q_S + Upsample(M_d \cdot Q_S) \tag{7}$$

Where *Upsample* indicates that the point cloud with low resolution is upsampled.

3.3 Loss function

3.3.1 Segmentation loss. The standard cross-entropy loss is used for the final segmentation results ($\mathcal{L}_{\text{SEG}}$) and the auxiliary segmentation head after the deep branch ($\mathcal{L}_{\text{AUX}}$).

$$\mathcal{L}(p_t) = -\frac{1}{N}(1 - p_t)^{\gamma} log(p_t) \tag{8}$$

Where N is the total number of classifications and γ on the $(1\text{-}p_t)$ parameter is the sample point weighing the sample weight, which is defined as the attenuation coefficient of the class. Where p_t is precision, the higher the precision, the smaller the loss value.

3.3.2 Super-resolution loss. Since the resolution of the point cloud through random sampling in the deep network is greatly reduced, in order to extract the segmentation loss

function of the deep branch layer separately, two ultra-point cloud losses are added to the deep branch layer in this paper. The corresponding *MLP* is used to extract the segmentation result of the deep branch layer in advance, assuming that the original point cloud is F_O. Conventional mean square error loss ($\mathcal{L}_{SR}$, $\mathcal{L}_{SD}$) was used to monitor the reconstructed point cloud F_{rec}.

$$\mathcal{L} = \|F_o - F_{rec}\|_2^2 \tag{9}$$

3.3.3 Overall loss. The overall loss $\mathcal{L}$ is a weighted combination of all above losses, which helps guide the model's learning and optimization more effectively:

$$\mathcal{L} = \mathcal{L}_{SEG} + \lambda_1 \mathcal{L}_{AUX} + \lambda_2 \mathcal{L}_{SR} + \lambda_3 \mathcal{L}_{SD} \tag{10}$$

Where λ_1 ,λ_2 ,λ_3 can help achieve a balance, ensuring that the model does not favor any specific task, thereby improving overall generalization capability.

4. Experiments

In this section, the S3DIS dataset (Armeni *et al*., 2017) and Sensaturban (Hu *et al*., 2022) are used to evaluate the performance of our proposed DSF-Netr for 3D semantic segmentation. This study reports the results of a comparison with current state-of-the-art methods, conducts ablation study and analyzes the effects of the proposed modules.

4.1 Experimental setup

(1) Training detail: The hardware environment of the experiment was Intel i7 12700K CPU and 112 GB memory. All experiments have conducted on NVIDIA RTX 3090 GPU. The initial learning rate was set to 0.01 and decreased by 5% after each epoch.

(2) Dataset: The S3DIS dataset is a collection of 3D point cloud data from 271 rooms in six regions, including 13 categories such as tables, chairs and ceilings. The SensatUrban (Hu *et al*., 2022) dataset contains 13 semantic classes, including major categories such as ground, buildings, transport roads and some minor categories such as bicycles, railways and Bridges.

(3) Metrics: Referring to PointNet++ (Zhao *et al*., 2022), RandLA-Net (Hu *et al*., 2020) and other methods, the evaluation indexes used in each experiment are mean intersection and merger ratio (MIOU), mean class accuracy (mAcc) and overall point-by-point accuracy (OA) to measure the model segmentation effect of DSF-Net.

4.2 Results

Table 1 respectively tests the total time consumption, point cloud input and memory consumption in semantic segmentation. By comparing the experimental data results of

Table 1. Performance comparison experiments on S3DIS

Methods	mIoU	Total time (seconds)	GPU memory (MB)
PointNet++ (Liu *et al*., 2019)	57.75	45,466	4,392
SPG (Zhao *et al*., 2022)	60.3	56,433	1,093
PointCNN (Que *et al*., 2021)	56.45	86,544	10,932
RandLA-Net (Hu *et al*., 2020)	51.57	19,326	1,563
DSF-Net(*ours*)	62.1	21,854	1,242

Note(s): SPG: super point graph
Source(s): Authors' own creation

different methods in the table, it can be observed that DSF-Net solves the problem of segmentation efficiency, achieves a balanced effect while ensuring the ability of a large range of point cloud input and reduces the memory consumption.

As shown in Table 2, the experimental results of our method and other comparsion models on the SensatUrban dataset for quantitative evaluation.

Table 2 shows that the DSF-Net framework achieved the best mIoU and mAcc among all methods, where mAcc is 1.11% higher than RandLA-Net and mIoU is 0.47% higher. According to the experiments in Table 1, it can be seen that the total time consumed by DSF-Net is slightly slower than that of RandLA-Net, but its efficiency is significantly better than other single-layer network structures. Compared with the existing models, the proposed model can achieve excellent results in most categories, mAcc and mIoU, and overall performance. To visualize the segmentation results, the semantic segmentation results for three different scenarios in SensatUrban are visualized and compared them with the results from PointNet and RandLA-Net. The visualization results are shown in Figure 4.

4.3 Ablation study

In order to verify the validity of each component in DSF-Net, an ablative study is conducted and tested it on the S3DIS dataset. It can be observed from Table 3 that for deep branch layer, when 1/4 scale point cloud is used as input, the segmentation accuracy is low and the segmentation efficiency is good. When the full-scale point cloud is used as input, the segmentation accuracy is higher, but the efficiency is lower. On the contrary, the experimental results of shallow branch layer are opposite to those of deep branch layer. For the baseline network combining deep and shallow branches, it achieves a good balance in segmentation efficiency and segmentation accuracy. At the 1/4 scale, the baseline network is twice as efficient as the deep branch layer. Besides, at full scale, the baseline network also achieves 65.15% accuracy compared to the shallow branch layer. Therefore, the network can get the best result between accuracy and efficiency.

To assess the effectiveness of the RAF module, a controlled variable approach uses a standard adder to replace the RAF module for experimental comparison. The experiments optimize the overall loss function presented in Section 3.4. As shown in Table 4, the RAF module achieves a satisfactory balance between accuracy and efficiency. In comparison with the standard adder, the RAF module demonstrates superior accuracy. Additionally, the RAF module requires less memory than the adder and exhibits faster inference speed in the experiments. Therefore, this module suits the separate fusion of feature vectors from both the deep and shallow branches.

5. Conclusion

In this paper, a point cloud semantic segmentation method based on DSF-Net is presented, and Deep training branch and shallow training branch are designed, respectively. Deep sampling is performed by random sampling method, and PASPP, based on void convolution, is used as input to enhance spatial details. Considering the relationship information between the integrated branches, a new feature fusion module RAF is introduced. The method in this paper improves the efficiency of semantic segmentation in large scenarios and achieves the best performance on two public datasets. The DSF-Net framework obtained the best mIoU and mAcc among all methods in the SensatUrban dataset, with mAcc 1.11% higher and mIoU 0.47% higher than RandLA-Net.

Table 2. SensatUrban dataset experiment

Method	OA (%)	mIoU (%)	mAcc (%)	(IoU %)class Ground	Veg-	Building	Wall	Bridge	Parking	Rail	Traffic	Street	Car	Footpath	Bike	Water
PointNet	80.78	22.75	23.71	67.96	89.52	80.05	0	0	3.95	0	31.55	0	35.14	0	0	0
PointNet++	84.3	35.06	32.92	72.46	94.24	84.77	2.82	2.09	25.79	0	31.54	11.42	38.84	7.12	0	56.93
SegCloud	85.27	37.29	37.29	69.93	94.55	88.87	32.83	12.58	15.77	*15.48*	30.63	22.96	56.42	0.54	0	44.24
SPG	88.66	40.93	42.66	74.10	97.9	94.2	63.3	7.5	24.2	0	30.10	34	74.4	0	0	54.8
RandLA-Net	*90.2*	56.43	57.58	87.1	*98.91*	95.33	74.4	28.69	41.38	0	55.99	*54.43*	85.67	*50.39*	0	71.30
DSF-Net(ours)	89.78	*56.9*	*58.69*	*89.11*	98.07	96.58	*88.40*	*50.45*	*61.62*	0	*66.67*	53.23	*86.14*	39.63	0	*71.31*

Note(s): SPG: super point graph
Source(s): Authors' own creation

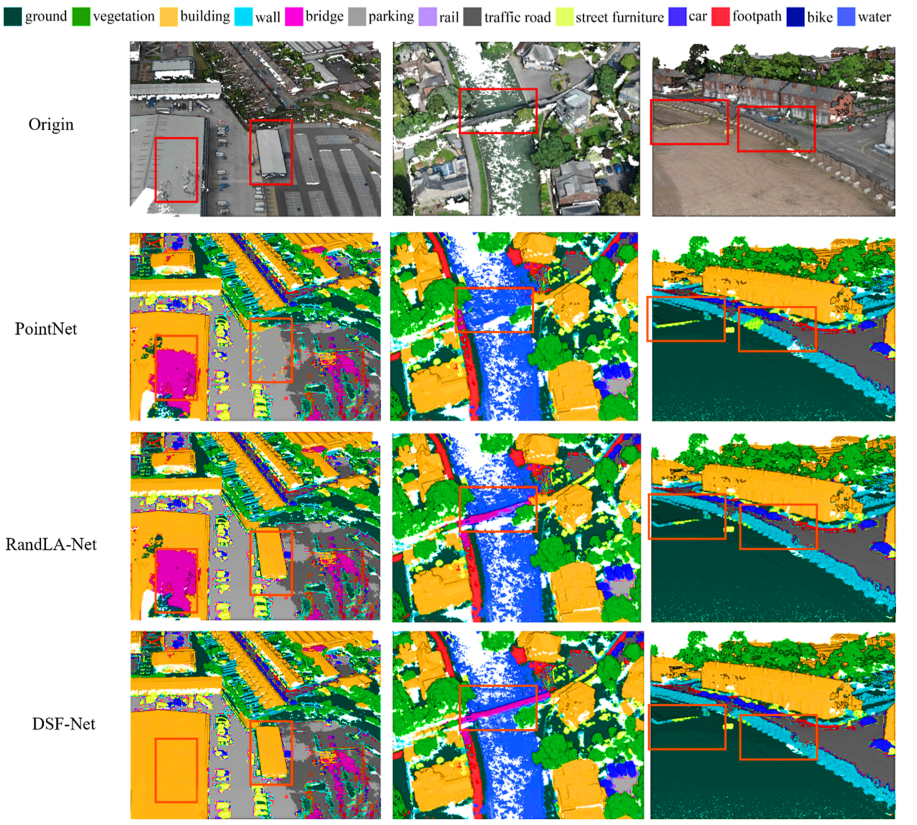

Figure 4. SensatUrban dataset visualization results. (Source: Authors' own creation)

Table 3. DSF-Net ablation study

D	S	Full scale	1/4 scale	mIoU	Time(s)
√			√	32.05	444,300
√		√		66.26	543,468
	√		√	46.15	112,354
	√	√		56.15	436,542
√	√		√	59.05	239,080
√	√	√		65.15	346,752

Source(s): Authors' own creation

Table 4. RAF ablation study

ADD	RAF	mIoU (%)	Total time (Seconds)
√		45.26	45,832
	√	62.15	23,548

Source(s): Authors' own creation

References

Armeni, I., Sax, S., Zamir, A.R. and Savarese, S. (2017), "Joint 2d-3d-semantic data for indoor scene understanding", *arxiv Preprint*, arxiv:1702.01105.

Chen, A., Xu, Z., Geiger, A., Yu, J. and Su, H. (2022), "Tensorf: tensorial radiance fields", *European Conference on Computer Vision*, Springer, pp. 333-350.

Croitoru, F.-A., Hondru, V., Ionescu, R.T., Shah, M. and Intelligence, M. (2023), "Diffusion models in vision: a survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45 No. 9, pp. 10850-10869, doi: 10.1109/tpami.2023.3261988.

Gao, Y., Liu, X., Li, J., Fang, Z., Jiang, X. and Huq, K.M.S. (2022), "LFT-Net: local feature transformer network for point clouds analysis", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 24 No. 2, pp. 2158-2168, doi: 10.1109/tits.2022.3140355.

Geng, X., Ji, S., Lu, M. and Zhao, L. (2021), "Multi-scale attentive aggregation for LiDAR point cloud segmentation", *Remote Sensing*, Vol. 13 No. 4, p. 691, doi: 10.3390/rs13040691.

Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K. (2003), "KNN model-based approach in classification", *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003*, Catania, Sicily, Italy, November 3-7, 2003, Springer, pp. 986-996, Proceedings.

Hamdi, A., Giancola, S. and Ghanem, B. (2021), "Voint cloud: multi-view point cloud representation for 3d understanding", *arxiv Preprint*, arxiv:2111.15363.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N. and Markham, A. (2020), "Randla-net: efficient semantic segmentation of large-scale point clouds", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11108-11117.

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N. and Markham, A. (2022), "Sensaturban: learning semantics from urban-scale photogrammetric point clouds", *International Journal of Computer Vision*, Vol. 130 No. 2, pp. 316-343, doi: 10.1007/s11263-021-01554-9.

Jhaldiyal, A. and Chaudhary, N. (2023), "Semantic segmentation of 3D LiDAR data using deep learning: a review of projection-based methods", *Applied Intelligence*, Vol. 53 No. 6, pp. 6844-6855, doi: 10.1007/s10489-022-03930-5.

Li, Y., Bu, R., Sun, M., Wu, W., Di, X. and Chen, B. (2018), "Pointcnn: convolution on x-transformed points", in *Advances in Neural Information Processing Systems*, Vol. 31.

Liu, Y., Fan, B., Meng, G., Lu, J., Xiang, S. and Pan, C. (2019), "Densepoint: learning densely contextual representation for efficient point cloud processing", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5239-5248.

Lyu, Y., Huang, X. and Zhang, Z. (2020), "Learning to segment 3d point clouds in 2d image space", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12255-12264.

Qi, C.R., Su, H., Mo, K. and Guibas, L.J. (2017a), "Pointnet: deep learning on point sets for 3d classification and segmentation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652-660.

Qi, C.R., Yi, L., Su, H. and Guibas, L.J. (2017b), "Pointnet++: deep hierarchical feature learning on point sets in a metric space", in *Advances in Neural Information Processing Systems*, Vol. 30.

Que, Z., Lu, G. and Xu, D. (2021), "Voxelcontext-net: an octree based framework for point cloud compression", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6042-6051.

Shuai, H., Xu, X. and Liu, Q. (2021), "Backward attentive fusing network with local aggregation classifier for 3D point cloud semantic segmentation", *IEEE Transactions on Image Processing*, Vol. 30, pp. 4973-4984, doi: 10.1109/tip.2021.3073660.

Su, Y., Liu, W., Yuan, Z., Cheng, M., Zhang, Z., Shen, X. and Wang, C. (2022), "DLA-Net: learning dual local attention features for semantic segmentation of large-scale building facade point clouds", *Pattern Recognition*, Vol. 123, 108372, doi: 10.1016/j.patcog.2021.108372.

Xu, W., Li, X., Ni, P., Guang, X., Luo, H. and Zhao, X. (2023), "Multiview fusion driven 3-D point cloud semantic segmentation based on hierarchical transformer", *IEEE Sensors Journal*, Vol. 23 No. 24, pp. 31461-31470, doi: 10.1109/jsen.2023.3328603.

Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J. and Lu, J. (2022), "Point-bert: pre-training 3d point cloud transformers with masked point modeling", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19313-19322.

Zeng, Z., Xu, Y., Xie, Z., Tang, W., Wan, J. and Wu, W. (2024), "Large-scale point cloud semantic segmentation via local perception and global descriptor vector", *Expert Systems with Applications*, Vol. 246, 123269, doi: 10.1016/j.eswa.2024.123269.

Zhang, S., Wang, B., Chen, Y., Zhang, S. and Zhang, W. (2024), "Point and voxel cross perception with lightweight CosFormer for large-scale point cloud semantic segmentation", *International Journal of Applied Earth Observation and Geoinformation*, Vol. 131, 103951, doi: 10.1016/j.jag.2024.103951.

Zhao, L., Xu, S., Liu, L., Ming, D. and Tao, W. (2022), "SVASeg: sparse voxel-based attention for 3D LiDAR point cloud semantic segmentation", *Remote Sensing*, Vol. 14 No. 18, 4471, doi: 10.3390/rs14184471.

Corresponding author
Jiawei Lu can be contacted at: viivan@cjlu.edu.cn