Pattern Recognition 154 (2024) 110630

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

TCFAP-Net: Transformer-based Cross-feature Fusion and Adaptive Perception Network for large-scale point cloud semantic segmentation

Jianjun Zhang [a,b,1], Zhipeng Jiang [b,1], Qinjun Qiu [a,1], Zheng Liu [a,c,*]

[a] School of Computer Science, China University of Geosciences (Wuhan), Wuhan, 430074, China
[b] National Engineering Research Center for Geographic Information System, China University of Geosciences (Wuhan), Wuhan, 430074, China
[c] Guangdong Enterprise Key Laboratory for Urban Sensing, Monitoring and Early Warning, Guangzhou, 510060, China

ARTICLE INFO

Keywords:
Transformer
Attention
Semantic segmentation
Point cloud scenarios

ABSTRACT

Point cloud semantic segmentation is an ingredient in understanding real-world scenes. Most existing approaches perform poorly on scene boundaries and struggle with recognizing objects of different scales. In this paper, we propose a novel framework that incorporates Transformer into the U-Net architecture for inferring pointwise semantics. Specifically, the Transformer-based cross-feature fusion module is designed first to employ geometric and semantic information to learn feature offsets to overcome the border ambiguity of segmentation results, and then it utilizes the Transformer to learn cross-feature enhanced and fused encoder features. Additionally, to facilitate the overall network's structure-to-detail perception capabilities, the adaptive perception module is designed, which employs cross-attention to adaptively allocate weights to encoder features at varying resolutions, establishing long-range contextual dependencies. Ablation studies validate the individual contributions of our module design choices. Compared with the existing competitive methods, our approach achieves state-of-the-art performance and exhibits superior results on benchmarks. Code is available at https://github.com/xiluo-cug/TCFAP-Net.

1. Introduction

Point clouds are a common data format used to depict real-world scenarios. Compared to 2D images, 3D point clouds have more significant geometric properties such as normal and curvature, allowing them to depict scene shapes and structures more accurately. Nowadays, point clouds can be easily acquired using depth sensors, LiDAR, and multi-view pictures, which are extensively utilized in a diverse spectrum of applications spanning from virtual reality, and environmental perception, to urban modeling, and robot simulation, owing to the rapid advancement of scanning techniques [1,2].

Semantic segmentation of large-scale point clouds is crucial for understanding real-world scenarios comprehensively. Aside from being utilized directly for scene understanding, point cloud semantics can effectively serve subsequent geometry processing tasks such as point cloud instance segmentation [3,4], point cloud registration [5,6], and surface reconstruction [7,8]. Compared with object-level and synthetic data, scene point clouds are often characterized by noise, large quantity, and irregular sampling. For the scene semantic labeling problem, deep learning-based approaches frequently outperform traditional methods in terms of robustness, accuracy, and generalization [9–11].

Effectiveness and efficiency are both critical for handling large-scale point cloud scenarios. PointNet [12], a pioneering work that directly applies neural networks on unstructured point clouds without auxiliary techniques. However, PointNet excels only in object-level point cloud labeling. In addition, the per-point operation or farthest-point sampling (FPS) strategy of the PointNet-based approaches also limits their computational efficiency. To trade off between effectiveness and efficiency on large-scale point clouds, RandLA-Net [9] employs random sampling (RS) as its sampling strategy, and proposes a feature aggregation module to preserve geometry structural information. Furthermore, RandLA-Net leverages the U-Net [13] structure to improve the network's ability to extract multiscale characteristics. Based on the encoder–decoder architecture, numerous efforts such as SCF-Net [14], BAAF-Net [15], and EyeNet [11] are proposed to improve and enhance the functionality of RandLA-Net. However, limitations and challenges still need to be addressed.

RandLA-Net and its variants [9,11,14,15] cannot accurately perceive contextual information at geometric boundaries, resulting in artifact misidentification. Furthermore, the traditional U-Net structure exhibits weak multiscale feature perception capability. For scene-level

* Correspondence to: No. 388 Lumo Road, Wuhan, China.
E-mail addresses: jianjunzhang@cug.edu.cn (J. Zhang), jiangzhipeng@cug.edu.cn (Z. Jiang), qiuqinjun@cug.edu.cn (Q. Qiu), liuzheng@cug.edu.cn (Z. Liu).
[1] Present/permanent address: No. 388 Lumo Road, Wuhan, China.

https://doi.org/10.1016/j.patcog.2024.110630
Received 10 December 2023; Received in revised form 24 March 2024; Accepted 23 May 2024
Available online 25 May 2024

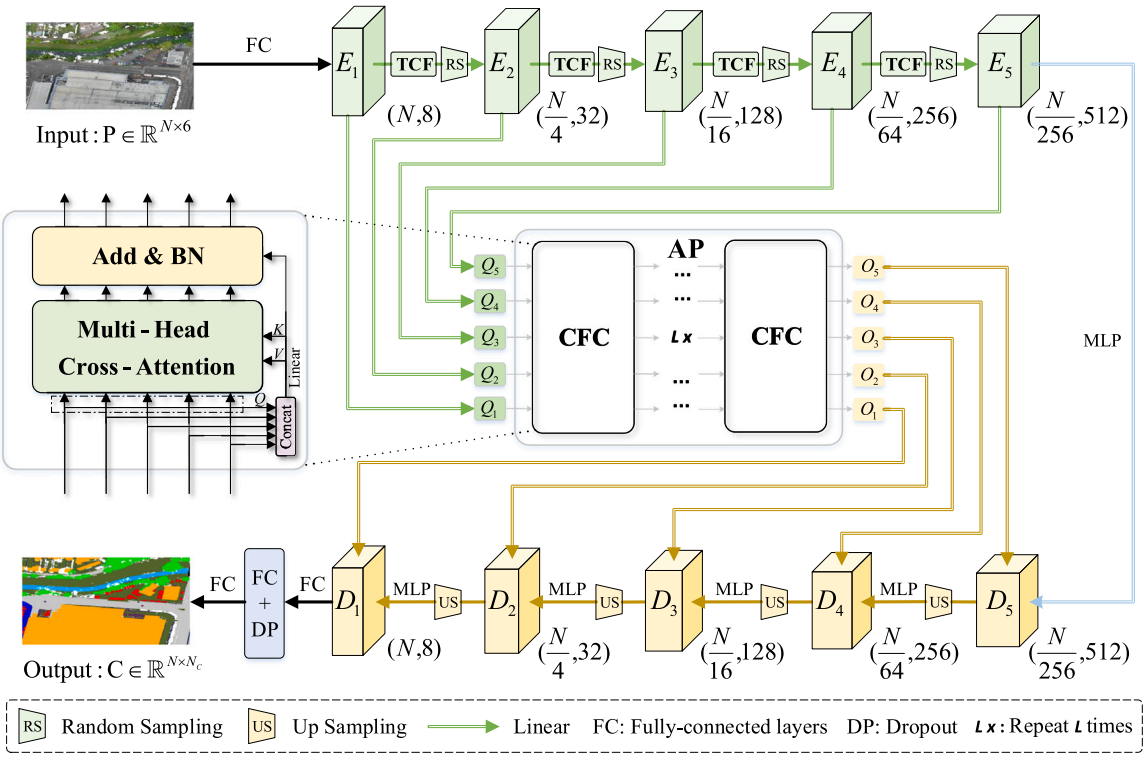

Fig. 1. The workflow of our proposed network TCFAP-Net for labeling large-scale scenarios. The core modules of TCFAP-Net are Transformer-based Cross-feature Fusion (TCF) and Adaptive Perception (AP).

point clouds, the structure-to-detail perception ability of multiscale features is crucial for producing high-quality segmentation results. Last but not least, the above approaches are primarily concerned with extracting and aggregating local features while ignoring long-range dependencies, making it difficult to capture long-range contextual information.

To overcome the deficiencies in the aforementioned methods, we propose a novel network framework dubbed TCFAP-Net. TCFAP-Net incorporates U-Net and Transformer into a unified framework and benefits from both. Our network's key modules are Transformer-based Cross-feature Fusion and Adaptive Perception, abbreviated TCF and AP. TCF consists of the Bilateral Feature Offset (BFO) unit and the Transformer Feature Fusion (TFF) unit, and AP comprises several cascaded Context Feature Coordination (CFC) units. On the one side, BFO learns the feature offset to clarify feature boundaries. Subsequently, TFF enhances and fuses geometric and semantic features to increase the saliency and consistency of the query points. On the other side, AP utilizes multi-head cross-attention to update encoder features at varying feature scales for establishing long-range dependencies, which can significantly improve our network's structure-to-detail perception capability. Our major contributions are summarized below:

- We propose a novel framework, coined as TCFAP-Net, that incorporates Transformer-based feature fusion and adaptive perception into the U-Net architecture for point cloud semantic segmentation.
- We present a TCF module to facilitate the fusion of geometric and semantic features, as well as local and global information, thereby distinguishing feature boundaries and improving the saliency and consistency of points.
- We provide an AP module comprised of cascaded Transformer-based blocks for adaptively perceiving multiscale features and effectively capturing long-range contextual information.
- Our approach achieves state-of-the-art performance and demonstrates highly competing results on several benchmarks.

2. Related work

2.1. Point cloud semantic segmentation

The success of deep learning in image processing [16–20] has led to the adoption of data-driven approaches for diverse point cloud processing tasks, including semantic segmentation. Based on different feature representations of point clouds, the deep learning methods of semantic segmentation can be categorized into two categories: direct and indirect ones. Specifically, the indirect methods typically rely on intermediate representations (such as voxel-based [21,22] and projection-based [23,24] representations) to bridge the gap between raw data and the desired deep-learning task. Compared to that, direct approaches do not incur information loss due to data transformation processes. PointNet [12] is a pioneering work that directly applies deep neural networks to point cloud classification and segmentation. PointNet++ [25], built upon PointNet, introduces a hierarchical processing structure via sampling and grouping operations to capture multi-scale information. Later on, PointCNN [26] introduces the X-Conv operator to aggregate the local information of point clouds, hence enabling appropriate characterization of spatial local correlations. Additionally, KPConv [27] conducts deformable convolution operations based on convolution kernels, which adapt to point clouds of varying sizes and densities.

To make a trade-off between performance and efficiency, Jing et al. [28,29] have made numerous outstanding contributions to 3D point cloud processing. For point cloud semantic segmentation, Hu et al. [9] proposed an encoder–decoder architecture, which incorporates random sampling and local feature aggregation to mitigate the loss of information during the down-sampling process. BAF-LAC [30] replaces the skip-connection in U-Net with a backward attentive fusion module, addressing the feature inconsistency between different coding layers. Qiu et al. [15] designed an adaptive multi-resolution feature fusion structure that effectively learns comprehensive knowledge of point clouds. SCF-Net [14] learns local context features of point clouds

by mapping them to the polar coordinate system. EyeNet [11] simulates the dual receptive fields of human vision by setting two different sizes of K-neighbors and processing the information from both receptive fields in parallel. Inspired by the above work, we also adopt the encoder–decoder structure and the random sampling strategy. However, we first incorporate the Transformer technique in the feature encoding and fusion phases.

2.2. Transformer and its variants

With the success of Transformer [31] in the field of natural language processing, Transformer and its variants have also demonstrated a certain dominance in the computer vision community. ViT [32] divides an image into multiple patches and inputs them as tokens into several transformer encoder blocks to extract features for image classification. Swin-Transformer [33] significantly reduces the computational cost and memory consumption by adopting a hierarchical design and patch merging operation for image processing tasks. Methods of [34–37] introduce different design approaches to capture long-range contextual information in 2D images. In recent years, Transformer has also shown considerable potential in 3D point cloud understanding. PCT [38] makes use of Transformer's positional insensitivity to cope with point clouds and use attention techniques to extract features from point clouds. Point Transformer [39], Point Transformer V2 [40] and Stratified Transformer [41] demonstrate that self-attention is well-suited for processing point cloud data since it is a basic point-set operation, which is order-invariant and quantity-invariant to the input. Inspired by the preceding work, we employ Transformer in local feature aggregation for extracting feature representations. Meanwhile, we replace skip connections in U-Net with Transformer blocks to establish long-range and global dependencies of point clouds.

3. Methodology

3.1. Workflow of TCFAP-Net

An overview of our network TCFAP-Net is illustrated in Fig. 1. Given a point cloud $\mathrm{P} = \{p_i\}_{i=1}^{N} \in \mathbb{R}^{N\times 6}$ with coordinate and color information as input, TCFAP-Net infers semantic labels $\mathrm{C} \in \mathbb{R}^{N\times N_C}$ of the point cloud as output, where N_C is the number of semantic classes.

The workflow of TCFAP-Net includes encoding, contextual perception, and decoding stages. In the encoding stage, we first use a fully connected (FC) layer to convert the input P into a high-dimensional latent feature F, and feed F into the encoder to produce the initial encoder feature E_1. Then, the encoder feature E_l of the current layer and P are fed into the TCF module, followed by random sampling, to produce the encoder feature E_{l+1} of the next layer. After cascaded encoding processes, five encoder features are obtained, denoted as $\{E_l\}_{l=1}^{5}$. In the contextual perception stage, we use the linear layers to transform the encoder features into query features $\{Q_l\}_{l=1}^{5}$, and then feed these features into the AP module composed of several CFC blocks for obtaining the corresponding connection features $\{O_l\}_{l=1}^{5}$. In the decoding stage, we feed the current layer's decoder feature D_l and the corresponding connection feature O_{l-1} into an upsampling operation and an MLP layer to produce the decoder feature of the next layer D_{l-1}. Finally, we progressively leverage three FC layers and a dropout (DP) to infer semantic labels C from the final decoder feature D_1. For clarity, we sketch the workflow of TCFAP-Net in Algorithm 1.

3.2. Transformer-based cross-feature fusion

The TCF module, composed of BFO and TFF units (shown in Fig. 2), enables the fusion of geometric and semantic features and local and global information.

Algorithm 1: TCFAP-Net Workflow

Input: point cloud $\mathrm{P} \in \mathbb{R}^{N\times 6}$
Output: predicted semantic labels $\mathrm{C} \in \mathbb{R}^{N\times N_C}$
initialization: $l = 1$
(1) Encoding stage
$\mathrm{F} \leftarrow \mathrm{FC(P)}$;
$E_1 \leftarrow \mathrm{FC(F)}$;
while $l \le 4$ **do**
  $\mathrm{F} \leftarrow E_l$;
  $E_{l+1} \leftarrow \mathrm{RS}(\mathbf{TCF}(\mathrm{P}, \mathrm{F}))$;
end
(2) Contextual perception stage
for *each* E_l **do**
  $Q_l \leftarrow \varphi(E_l)$;
end
obtain: $\{O_1, O_2, O_3, O_4, O_5\} \leftarrow \mathbf{AP}(Q_1, Q_2, Q_3, Q_4, Q_5)$;
(3) Decoding stage
$D_5 \leftarrow \mathrm{TransConv}(\mathrm{MLP}(E_5) \oplus O_5),\ l = 5$;
while $l > 1$ **do**
  $D_{l-1} \leftarrow \mathrm{TransConv}(\mathrm{US}(D_l) \oplus O_{l-1})$;
end
(4) Prediction
$\mathrm{C} \leftarrow \mathrm{FC}(\mathrm{DP}(\mathrm{FC}(\mathrm{FC}(D_1))))$;
return C.

3.2.1. Bilateral feature offset

Given the point cloud P and the learned semantic feature F, the BFO unit first employs KNN searching to establish the neighborhood information graph and obtain the neighborhood information $P_i^K = \{p_i^1, p_i^2, \ldots, p_i^K\}$ of each point p_i and the corresponding semantic feature $F_i^K = \{f_i^1, f_i^2, \ldots, f_i^K\}$. For each point p_i, we perform relative position encoding (RPE) and relative semantic encoding (RSE) to learn local geometric and semantic features G_{p_i} and S_{f_i}, respectively

$$G_{p_i} = \mathrm{MLP}\left(\overline{P_i} \oplus \left(\overline{P_i} - P_i^K\right)\right), \tag{1}$$

$$S_{f_i} = \mathrm{MLP}\left(\overline{F_i} \oplus \left(\overline{F_i} - F_i^K\right)\right), \tag{2}$$

where $\oplus$ is the concatenation operation, $\overline{P_i}$ is the broadcast of p_i (i.e., copy itself to align the number dimension to K), and $\overline{F_i}$ is the broadcast of f_i.

The direct concatenation often leads to feature ambiguity, especially when diverse semantic items are present in a local perception space; see the boundary regions in Fig. 3 for example. To overcome this ambiguity, the BFO unit is presented, which interactively shifts geometric and semantic features according to the feature offsets acquired from each other. For the geometric feature G_{p_i} of p_i, BFO can learn pull-in offset from the semantic feature S_{f_i} to bring features of the same semantic class of p_i closer together, as well as pull-out offset to push features of different semantic classes farther apart. For the semantic feature S_{f_i}, BFO can learn offset from the geometric feature G_{p_i} similarly. Specifically, bilateral offset features $\tilde{G}_{p_i}$ and $\tilde{S}_{f_i}$ can be obtained by learning shifting offsets as follows:

$$\tilde{G}_{p_i} = G_{p_i} \oplus \left(\mathrm{MLP}\left(S_{f_i}\right) + G_{p_i}\right), \tag{3}$$

$$\tilde{S}_{f_i} = S_{f_i} \oplus \left(\mathrm{MLP}\left(G_{p_i}\right) + S_{f_i}\right). \tag{4}$$

Fig. 3 exhibits the above-mentioned bilateral offset learning process.

3.2.2. Transformer-based feature fusion

Bilateral (geometric and semantic) offset features $\tilde{G}_{p_i}$ and $\tilde{S}_{f_i}$ are obtained through the proposed BFO unit. To further enrich the feature representations, we propose the TFF unit which utilizes self- and cross-attention mechanisms to emphasize and fuse bilateral features, as illustrated in Fig. 2.

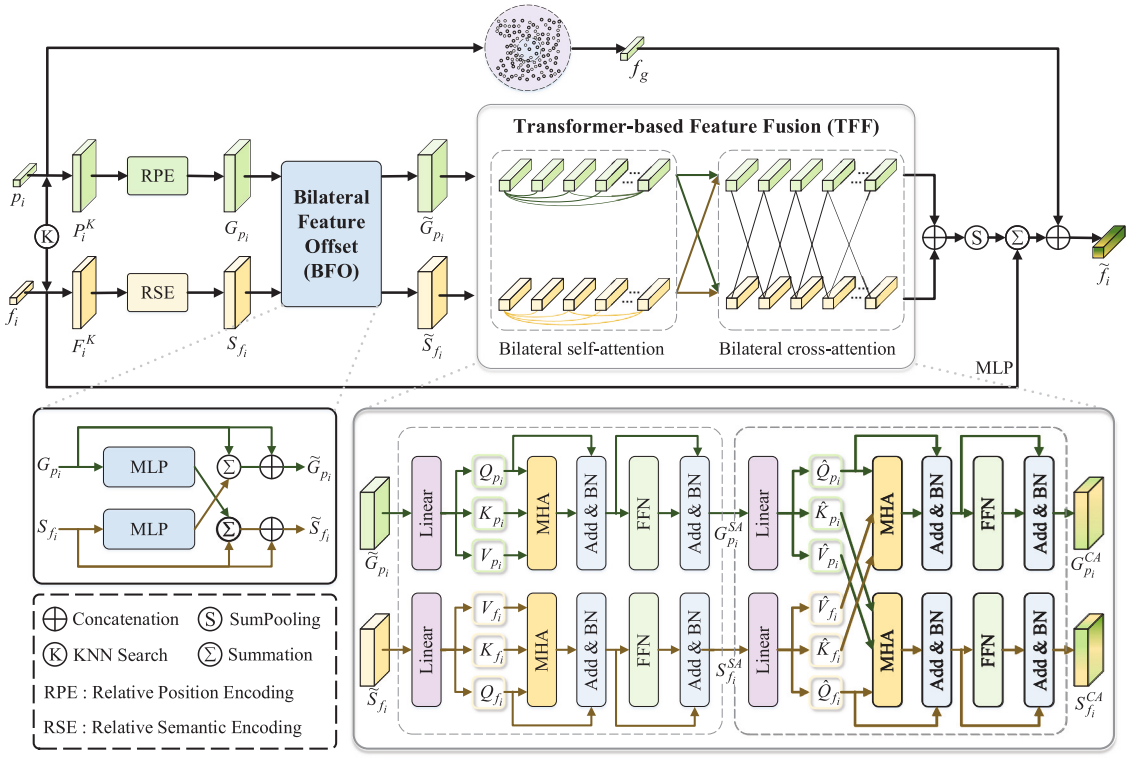

Fig. 2. Illustration of our proposed Transformer-based Cross-feature Fusion (TCF) module incorporating Bilateral Feature Offset (BFO) and Transformer-based Feature Fusion (TFF) units. The top panel is the encoding workflow for producing the encoder feature. The bottom panel illustrates the design details of BFO and TFF units.

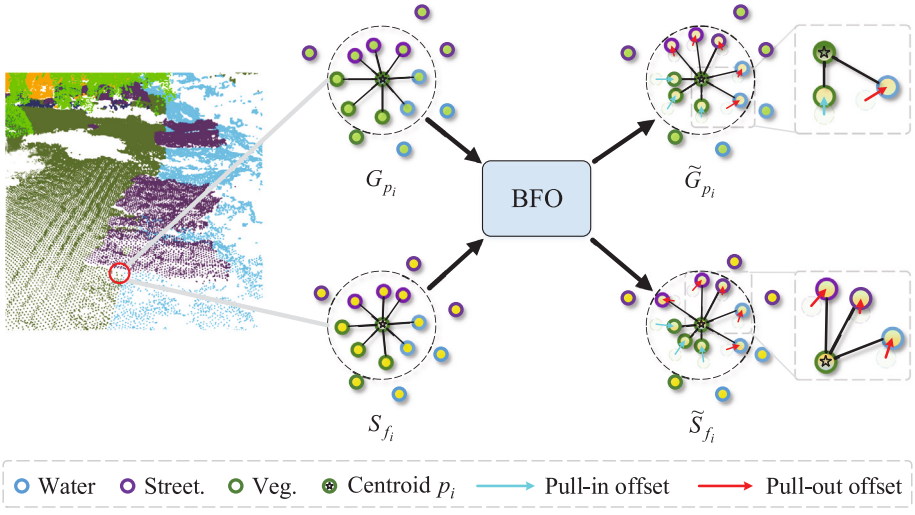

Fig. 3. Illustration of the learning process for bilateral offset features $\tilde{G}_{p_i}$ and $\tilde{S}_{f_i}$, which pull-in similar neighborhood features together and pull-out dissimilar features farther away.

Function Definitions in Transformer. The primary mechanism of the Transformer is attention. Given the feature token F_{in} as input, three weight matrices W_Q, W_K, and W_V can be learned by linear projections, which map the input to three separate feature spaces to obtain the query Q, key K, and value V as $\{Q, K, V\} = \{F_{in}W_Q, F_{in}W_K, F_{in}W_V\}$. Then, the attention function is formulated as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{5}$$

where d_k is the dimension of K. The function (5) first computes the weight matrix by using the pair of query Q and key K, and then utilizes the weight matrix to update the value V.

However, the aforementioned single attention (5) struggles to capture information from varying feature spaces. Thus, multi-head attention, abbreviated as MHA, is introduced.[31]. Specifically, MHA performs attention function (5) h times to produce multiple attention heads within different feature spaces, which are subsequently concatenated as follows:

$$\begin{aligned} &\text{MHA}(Q, K, V) = \text{HA}_1 \oplus \text{HA}_2 \oplus \cdots \oplus \text{HA}_h; \\ &\text{HA}_h = \text{Attention}(Q_h, K_h, V_h). \end{aligned} \tag{6}$$

We next apply residual connections and normalization in MHA (6) to improve the efficiency and stability of the overall network as

$$\psi(Q, K, V) = \text{BN}(\text{MHA}(Q, K, V)) + Q, \tag{7}$$

where BN is batch normalization. In order to further capture complex nonlinear information of point clouds, a feed-forward network, abbreviated as FFN, is employed to refine the result produced by (7). Therefore, the final attention function $\mathcal{F}(\cdot)$ in TFF can be defined as

$$\mathcal{F}(Q,K,V) = \mathrm{BN}(\mathrm{FFN}(\psi(Q,K,V))) + \psi(Q,K,V). \tag{8}$$

Bilateral self-attention. To capture the correlations between neighborhoods and obtain a more comprehensive context, we use the self-attention mechanism to dynamically adjust the feature weights of each neighboring point, aiming to achieve better local feature representation.

In particular, we apply linear projections on the features $\tilde{G}_{p_i}$ and $\tilde{S}_{f_i}$, respectively, the linear projection operation is defined as $\varphi(\cdot)$, to produce two sets of query, key, and value as follows:

$$\{Q_{p_i}, K_{p_i}, V_{p_i}\} = \{\varphi(\tilde{G}_{p_i}), \varphi(\tilde{G}_{p_i}), \varphi(\tilde{G}_{p_i})\}, \tag{9}$$

$$\{Q_{f_i}, K_{f_i}, V_{f_i}\} = \{\varphi(\tilde{S}_{f_i}), \varphi(\tilde{S}_{f_i}), \varphi(\tilde{S}_{f_i})\}. \tag{10}$$

Then the $\mathcal{F}(\cdot)$ (8) is employed to produce the below self-attention enhanced features:

$$G^{SA}_{p_i} = \mathcal{F}\left(Q_{p_i}, K_{p_i}, V_{p_i}\right), \tag{11}$$

$$S^{SA}_{f_i} = \mathcal{F}\left(Q_{f_i}, K_{f_i}, V_{f_i}\right). \tag{12}$$

As illustrated in Fig. 2, we finally acquire enhanced features $G^{SA}_{p_i}$ and $S^{SA}_{f_i}$ from the bilateral offset features $\tilde{G}_{p_i}$ and $\tilde{S}_{f_i}$, respectively.

Bilateral cross-attention. To facilitate mutual guidance between geometric and semantic features, we design a bilateral cross-attention unit to establish connections between self-attention enhanced features $G^{SA}_{p_i}$ and $S^{SA}_{f_i}$ and ensure their inherent fusion.

Specifically, we first perform linear projections on the geometric and semantic features to obtain two sets of query, key, value features, which are as follows:

$$\{\hat{Q}_{p_i}, \hat{K}_{f_i}, \hat{V}_{f_i}\} = \{\varphi(G^{SA}_{p_i}), \varphi(S^{SA}_{f_i}), \varphi(S^{SA}_{f_i})\}, \tag{13}$$

$$\{\hat{Q}_{f_i}, \hat{K}_{p_i}, \hat{V}_{p_i}\} = \{\varphi(S^{SA}_{f_i}), \varphi(G^{SA}_{p_i}), \varphi(G^{SA}_{p_i})\}. \tag{14}$$

Then, we feed the two sets of query, key, and value features (13) and (14) into $\mathcal{F}(\cdot)$ (8), respectively, to obtain the cross-attention fused features as follows:

$$G^{CA}_{p_i} = \mathcal{F}\left(\hat{Q}_{p_i}, \hat{K}_{f_i}, \hat{V}_{f_i}\right), \tag{15}$$

$$S^{CA}_{f_i} = \mathcal{F}\left(\hat{Q}_{f_i}, \hat{K}_{p_i}, \hat{V}_{p_i}\right). \tag{16}$$

As shown in Fig. 2, our method achieves the fusion of geometric and semantic features, with the former providing local position embedding and the latter supplying comprehensive contextual information.

Cross-feature information fusion. Inspired by the work [14], we incorporate the volume ratio feature f_g into the encoding procedure for effectively preserving global structural properties In addition, f_i serves as a residual and is added to the fused features to prevent the loss of detail. The above process is formulated as

$$\tilde{f}_i = \left(\mathrm{SumPooling}\left(G^{CA}_{p_i} \oplus S^{CA}_{f_i}\right) + \mathrm{MLP}\left(f_i\right)\right) \oplus f_g. \tag{17}$$

Thus, we can learn the pointwise semantic feature $\tilde{f}_i$ and finally derive encoder feature as $E_l = \{\tilde{f}_i | \forall p_i \in \mathrm{P}\}$.!

In summary, the learned encoder feature E_l has a comprehensive representation in both geometric and semantic feature space, including refined local information and the global property, thereby improving the accuracy and robustness of semantic segmentation.

3.3. Adaptive perception

The well-known U-Net architecture [13] leverages skip connections to concatenate encoder and decoder features from the same layer to compensate for information loss during the decoding phase. However, without mutual support and compensation between features at different scales, U-Net falls short in its multiscale feature perception capabilities, affecting its ability to gather contextual information from the entire point cloud. To tackle the problem, we develop the Adaptive Perception (AP) module, which employs multi-head cross-attention [31], allowing the encoder feature of each layer to adaptively perceive multiscale semantics and capture long-range contextual information within the entire point cloud.

In particular, AP consists of L stacked Context Feature Coordination (CFC) units. We first obtain query features $\{Q_l\}_{l=1}^5$ from all encoder features $\{E_l\}_{l=1}^5$ using a linear layer as

$$Q_l = \varphi(E_l), \quad l = 1,2,3,4,5. \tag{18}$$

Then, the query features $\{Q_l\}_{l=1}^5$ are fed into CFC units to produce connection features $\{O_l\}_{l=1}^5$. We sample the query features to the resolution of the lth encoder using the alignment function! $Y(\cdot)$,

$$Y(Q_j) = \begin{cases} \mathrm{RS}_{j\to l}(Q_j), & \text{if } j < l, \\ Q_j, & \text{if } j = l, \\ \mathrm{US}_{j\to l}(Q_j), & \text{if } j > l; \end{cases} \tag{19}$$

then concatenate the aligned query features to produce Q_{Σ_l}, followed by creating key K_l and value V_l using the linear layer,

$$Q_{\Sigma_l} = Y(Q_1) \oplus Y(Q_2) \oplus \cdots \oplus Y(Q_5), \tag{20}$$

$$\left\{K_l, V_l\right\} = \left\{\varphi(Q_{\Sigma_l}), \varphi(Q_{\Sigma_l})\right\}. \tag{21}$$

The key and value pair $\{K_l, V_l\}$ can serve as contextual information to update the query feature Q_l for each layer, as shown in Fig. 4. Multi-head cross-attention enables each layer's query feature to capture the long-term dependency of contextual information, strengthening highly relevant characteristics of query features while suppressing irrelevant information. Thus, we utilize $\psi(\cdot)$ (7) to learn the connection features $\{O_l\}_{l=1}^5$ as

$$O_l = \psi\left(Q_l, K_l, V_l\right), \quad l = 1,2,3,4,5. \tag{22}$$

At last, decoder features $\{D_l\}_{l=1}^5$ are obtained as follows:

$$D_{l-1} = \begin{cases} \mathrm{TransConv}(\mathrm{MLP}(E_{l-1}) \oplus O_{l-1}), & l = 6, \\ \mathrm{TransConv}(\mathrm{US}(D_l) \oplus O_{l-1}), & l = 2,3,4,5, \end{cases} \tag{23}$$

where TransConv is the transposed convolution.

As a result, the final decoding features are highly discriminative, context-aware features that can effectively express geometric-semantic information relevant to adaptive perception, assisting the classifier in identifying the same semantic categories and discriminating among different ones, thereby improving semantic segmentation performance.

4. Experiments and analysis

In this section, comprehensive experiments are designed to validate the superiority of TCFAP-Net across diverse benchmark datasets. First, the experimental setups are introduced in Section 4.1. Then, the benchmark description is presented in Section 4.2, and the comparison experiments are designed in Section 4.3. Finally, the ablation studies are conducted in Section 4.4, and the discussions are introduced in Section 4.5.

4.1. Experimental setups

To quantify the performance, overall accuracy (OA) and mean intersection over union (mIoU) are employed as criteria to evaluate performance. We adopt multiple cross-entropy loss as our energy function. Referring to the parameter configuration of [9], we set the number of neighbors for KNN searching to 16. The initial learning rate to 0.01, and the decay rate to 95%. Also, we conduct training for 200 epochs on indoor scenarios and 400 epochs on outdoor scenarios. The number

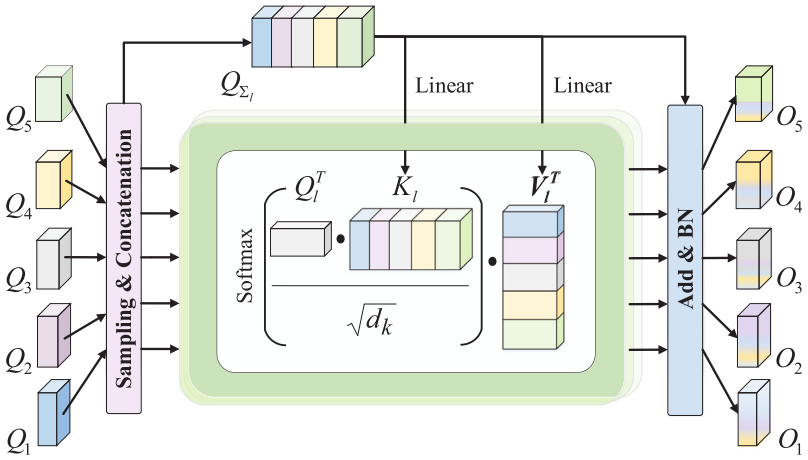

Fig. 4. Illustration of the Context Feature Coordination (CFC) unit within the AP module, which employs multi-head attention to adaptively perceive long-range contextual information.

Table 1
Numerical results of our approach and the compared methods on SensatUrban (%). The best and second-best results are highlighted in bold and underlined, respectively.

Method	Year	OA	mIoU	IoUs												
				Ground	Veg.	Build.	Wall	Bridge	Park.	Rail	Traffic.	Street.	Car	Foot.	Bike	Water
PointNet [12]	2017	80.8	23.7	67.9	89.5	80.1	0.0	0.0	3.9	0.0	31.6	0.0	35.1	0.0	0.0	0.0
PointNet++ [25]	2017	84.3	32.9	72.5	94.2	84.8	2.7	2.1	25.8	0.0	31.5	11.4	38.8	7.1	0.0	56.9
TagentConv [43]	2018	76.9	33.3	71.5	91.4	75.9	35.2	0.0	45.3	0.0	26.7	19.2	67.6	0.0	0.0	0.0
SPGraph [44]	2018	85.3	37.3	69.9	94.6	88.9	32.8	12.6	15.8	15.5	30.6	22.9	56.4	0.5	0.0	44.2
SparseConv [45]	2018	88.7	42.7	74.1	97.9	94.2	63.3	7.5	24.2	0.0	30.1	34.0	74.4	0.0	0.0	54.8
KPConv [27]	2019	93.2	57.6	**87.1**	**98.9**	95.3	**74.4**	28.7	41.4	0.0	55.9	**54.4**	**85.7**	40.4	0.0	**86.3**
RandLA-Net [9]	2020	89.8	52.7	80.0	98.1	91.6	48.9	40.6	51.6	0.0	56.7	33.2	80.0	32.6	0.0	71.3
BAF-LAC [30]	2021	91.5	54.1	84.4	98.4	94.1	57.2	27.6	42.5	15.0	51.6	39.5	78.1	40.1	0.0	75.2
BAAF-Net [15]	2021	91.8	56.1	83.3	98.2	94.0	54.2	51.0	57.0	0.0	60.4	14.0	81.3	41.6	0.0	58.0
NeiEA-Net [46]	2023	91.7	57.0	83.3	98.1	93.4	50.1	61.3	57.8	0.0	60.0	41.6	82.4	42.1	0.0	71.0
Eye-Net [11]	2023	**93.7**	62.3	86.6	98.6	**96.2**	65.8	59.2	**64.7**	**17.9**	**64.8**	49.8	83.1	**46.2**	11.1	65.4
TCFAP-Net	2023	92.6	**64.1**	85.7	98.6	95.2	60.2	**68.1**	63.1	16.4	56.1	48.3	83.4	42.6	**28.7**	78.9

of points for outdoor scenarios is restricted to 65,536, while the number for indoor scenarios is limited to 40,960. All experiments are conducted on one NVIDIA A100 GPU (80G), and the software environment is Ubuntu 20.04 with the TensorFlow deep learning framework installed.

4.2. Benchmark description

The benchmark datasets utilized in our experiments can be divided into three categories: urban, indoor, and street-view. For each benchmark dataset, we employ the position and color information of point clouds as inputs.

SensatUrban. SensatUrban [42] is an urban-scale point cloud dataset generated by UAV-based photogrammetry. The dataset spans multiple cities in the United Kingdom, covering an area of about 7.6 square kilometers and having roughly 3 billion points with rich semantic labels. The labeled semantics can be divided into 13 categories: ground, vegetation, building, car, railway, bicycle, etc. SensatUrban is separated into 34 tiles to generate training, validation, and test sets. For fairness of comparisons, we follow the official partitioning strategy for training and testing our approach and the competing methods.

S3DIS. S3DIS [47] dataset is a prominent indoor scene dataset acquired by Matterport scanners. S3DIS supports both semantic segmentation annotation as well as instance segmentation validation. The dataset is partitioned into six areas, with a total of 272 rooms covering an area of about 6000 square meters. It includes 13 indoor semantic objects, such as ceiling, floor, wall, window, and others. In this work, we select Area 5 for testing our approach and the compared ones while utilizing Areas 1–4 and 6 for training.

Toronto3D. Toronto3D [48] is a classic street-view dataset collected on a one-kilometer-long street in Toronto, Canada, using vehicle-mounted LiDAR technology. The dataset contains 78.3 million points with semantic information classified into eight categories, such as road, building, automobile, etc, making it excellent for portraying city scenarios. The dataset is partitioned into four blocks, with L002 serving as the test set and the remaining as the train set. The ratio of points in the test set to those in the training set is around 1:7. Similarly, we adopt the above official strategy to train and test our approach and the competing methods.

4.3. Comparison experiments

In this subsection, we perform numerical and visual comparisons on SensatUrban, S3DIS, and Toronto3D datasets to verify the superiority of our proposed TCFAP-Net over the compared competing approaches.

Evaluation on SensatUrban. Table 1 presents a quantitative comparison between TCFAP-Net and other competing methods. TCFAP-Net, in particular, exhibits a remarkable improvement of around 12% over the baseline [9], as well a considerable 2% improvement over the latest powerful approach EyeNet [11], showing that our method achieves state-of-the-art (SOTA) performance. Fig. 5 presents a visual comparison of TCFAP-Net and RandLA-Net. The above table and figure demonstrate that TCFAP-Net is capable of effectively segmenting objects at different scales, thanks to its powerful multiscale perception capabilities. Fig. 6 implies that TCFAP-Net can obtain more discriminative features than BAAF-Net.

Evaluation on S3DIS. Table 2 shows the numerical results of our approach and other competitive methods on S3DIS (Area 5). Notably, our approach produces the highest OA and mIoU values compared to the other methods. In addition, our approach has attained a leading position in the categories of chairs, boards, and clutters. Fig. 7 visualizes the segmentation results of our method and the baseline method. We

Fig. 5. Visual comparison of RandLA-Net (baseline) with our approach TCFAP-Net on SensatUrban.

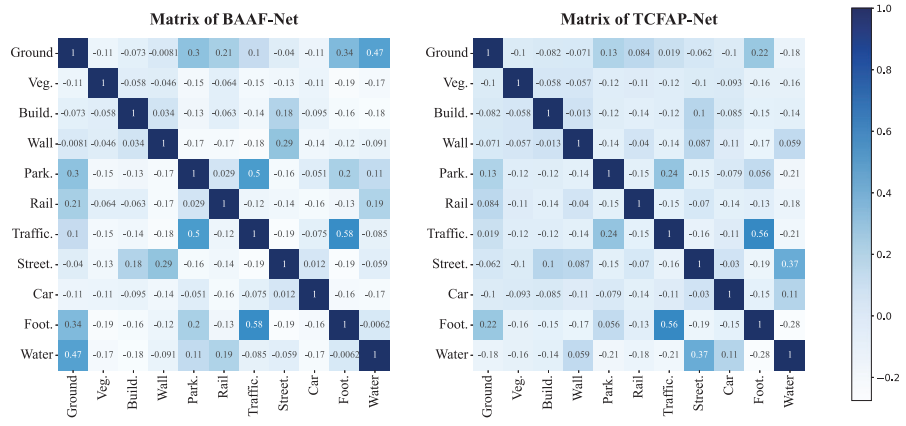

Fig. 6. The correlation matrices of high-dimensional features for 11 categories in the Birmingham Block 1 of Sensaturban, produced by BAAF-Net [15] and TCFAP-Net. The average correlation value of BAAF-Net is 0.071 and that of TCFAP-Net is 0.036, which shows that our approach can learn more discriminative features than BAAF-Net.

observe that our method produces superior segmentation results with clearer boundaries in regions of high similarity (e.g., boards and walls) and in category-dense regions (e.g., doors, floors, sofas, and walls). In contrast, RandLA-Net yields visually blurring results on the regions of boards, doors, floors, etc. To further evaluate the performance of our method across the entire scene of S3DIS, we conduct experiments using a sixfold cross-validation. First, we compare TCFAP-Net, RandLA-Net, and KPConv [27] in Fig. 8. Obviously, our TCFAP-Net obtains significantly higher segmentation accuracy in Area1, Area4, and Area5 compared to the other two approaches. Then, we present the numerical results of our approach and the compared methods in Table 3. Our approach obtains mIoU and OA of 72.5% and 89.3%, marking an improvement of 2.5% and 1.3%, respectively, over RandLA-Net (baseline). Our approach outperforms Point Transformer [39] and Stratified Transformer [41] in terms of model complexity and efficiency, despite lower segmentation accuracy on S3DIS, as detailed in the model complexity and efficiency part of Section 4.5. Thus, our approach strikes a balance between model effectiveness, complexity, and efficiency on S3DIS.

Evaluation on Toronto3D. Unlike urban environments, city street scenes are characterized by more additional details and infrastructure elements. To verify the effectiveness of TCFAP-Net for street scene data, the experiments are designed on the Toronto3D dataset.

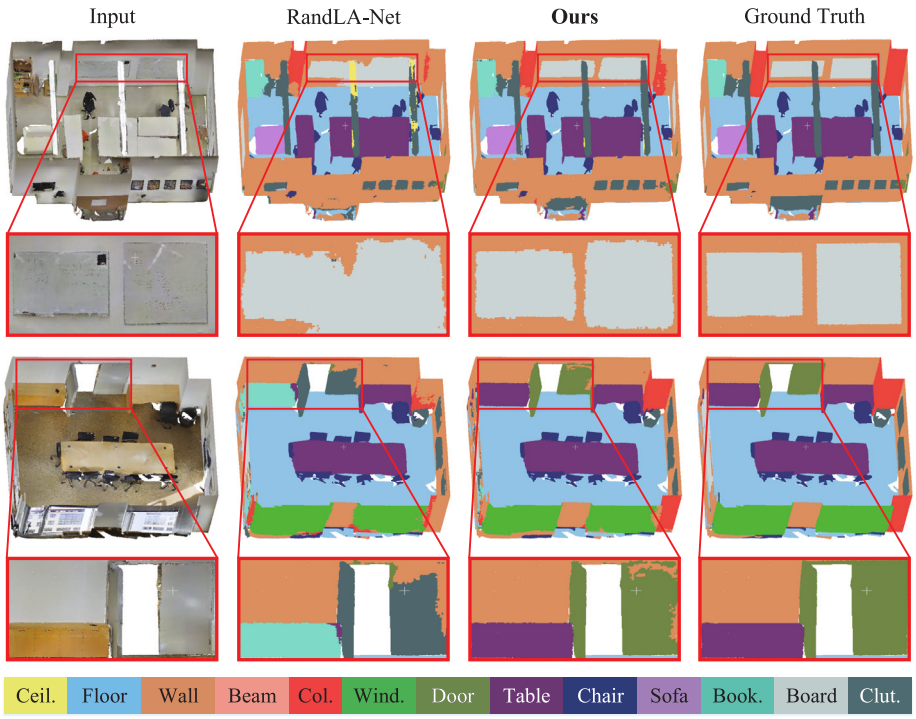

Fig. 7. Visual comparison of RandLA-Net (baseline) with our approach TCFAP-Net on S3DIS.

Table 2
Numerical results of our approach and the compared methods on S3DIS (Area5) (%). The best and second-best results are highlighted in bold and underlined, respectively.

Method	Year	OA	mIoU	IoUs												
				Ceil.	Floor	Wall	Beam	Col.	Wind.	Door	Table	Chair	Sofa	Book.	Board	Clut.
PointNet [12]	2017	–	41.1	88.8	97.3	69.8	0.1	3.9	46.3	10.8	59.0	52.6	5.9	40.3	26.4	33.2
TangentConv [43]	2018	–	52.6	90.5	97.7	74.0	0.0	20.7	39.0	31.3	77.5	69.4	57.3	38.5	48.8	39.8
PointCNN [26]	2018	85.9	57.3	92.3	98.2	79.4	0.0	17.6	28.8	62.1	70.4	80.6	39.7	66.7	62.1	56.7
SPGraph [44]	2018	86.4	58.0	89.4	96.9	78.1	0.0	**42.8**	48.9	61.6	**84.7**	75.4	69.8	52.6	2.1	52.5
RandLA-Net [9]	2020	87.2	62.4	91.1	95.6	80.2	0.0	24.7	62.3	47.7	76.2	83.7	60.2	71.1	65.7	53.8
BAAF-Net [15]	2021	**88.9**	65.4	92.9	97.9	82.3	0.0	23.1	**65.5**	**64.9**	78.5	87.5	61.4	70.7	68.7	57.2
BAF-LAC [30]	2021	–	65.7	91.9	97.4	82.0	0.0	19.9	61.5	52.9	80.3	87.8	**78.9**	**72.7**	75.0	53.8
LGGCM [49]	2022	88.8	63.3	**94.8**	**98.3**	81.5	0.0	35.9	63.3	43.5	80.2	88.4	68.8	55.8	64.6	47.8
NeiEA-Net [46]	2023	88.5	**66.1**	92.9	97.4	**83.3**	0.0	34.9	61.8	53.3	78.8	86.7	77.1	69.5	67.9	54.2
TCFAP-Net	2023	**88.9**	**66.1**	93.4	97.8	83.1	0.0	31.2	64.7	45.4	79.9	**89.2**	74.1	72.5	**76.8**	**57.6**

Table 3
Numerical results of our approach and the compared methods on S3DIS (6-fold). The best and second-best results are highlighted in bold and underlined, respectively.

Method	Year	mIoU (%)	OA (%)	mAcc (%)
PointNet [12]	2017	47.6	78.6	66.2
PointNet++ [25]	2017	64.5	81.0	67.1
KPConv [27]	2019	70.6	–	79.1
RandLA-Net [9]	2020	70.0	88.0	82.0
SCF-Net [14]	2021	71.6	88.4	82.7
BAAF-Net [15]	2021	72.2	88.9	**83.1**
BAF-LAC [30]	2021	71.7	88.2	81.3
Point Transformer [39]	2021	73.5	90.2	81.9
Stratified Transformer [41]	2022	**73.7**	**90.8**	81.7
TCFAP-Net	2023	72.5	89.3	81.0

As demonstrated in Table 4, our approach is comparable to most methods on mIoU and outperforms them in OA, especially for nature and building categories. Fig. 9 shows the visual comparison of semantic predictions in block L002. Our method outperforms RandLA-Net in terms of accuracy, as seen from the zoomed-in views in Fig. 9. This visual experiment demonstrates the fine segmentation capability of our method in large-scale street scenarios.

4.4. Ablation studies

In this section, for fair comparisons, all the tested models follow the same official training and testing strategies aforementioned in Section 4.2 on the S3DIS dataset.

4.4.1. Ablation study of the core modules

To assess the contribution of each core module in our TCFAP-Net, we perform separate ablated experiments for the BFO unit, the TFF unit, and the AP module. All the tested models are listed as follows:

Baseline. RandLA-Net [9] is chosen as the baseline because of its effectiveness in semantic labeling on scene point clouds.

A1. We replace the Local Spatial Encoding (LocSE) module of RandLA-Net with our BFO unit.

A2. We reconstruct the encoder in RandLA-Net with our BFO and TFF units.

Full model. The full model consists of our proposed BFO, TFF, and AP modules.

Observed from the ablation result, shown in Table 5, the full model performs best overall, showing that each core module applied in our full model improves its performance. The BFO unit achieved an improvement of 1.0% in mIoU compared to the baseline, while the TFF module

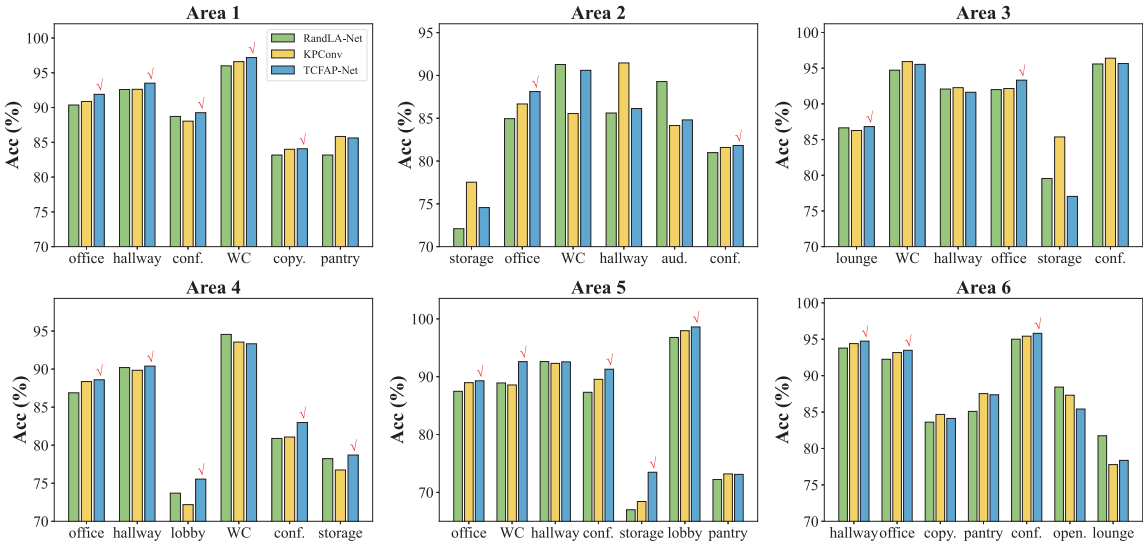

Fig. 8. Accuracy histograms of RandLA-Net, KPConv, and our TCFAP-Net on Area 1–6 of S3DIS. The x-axis represents subregions within each area, and the y-axis represents the accuracy (Acc) of those subregions. The symbol "√" denotes that our method outperforms RandLA-Net and KPConv in the corresponding subregion.

Table 4
Numerical results of our approach and the compared methods on Toronto3D (%). The best and second-best results are highlighted in bold and underlined, respectively.

Method	Year	OA	mIoU	IoUs							
				Road	Rmrk.	Natural	Build.	Util. line	Pole	Car	Fence
RandLA-Net [9]	2020	94.4	81.8	96.7	64.2	96.9	94.2	88.0	77.8	93.4	42.9
ResDLPS-Net [50]	2021	96.5	80.3	95.8	59.8	96.1	90.9	86.8	79.9	89.4	43.3
BAAF-Net [15]	2021	94.2	81.2	96.8	67.3	96.8	92.2	86.8	82.3	93.1	34.0
BAF-LAC [30]	2021	95.2	**82.0**	96.6	64.7	96.4	91.6	86.1	83.9	93.2	**43.5**
Point Transformer [39]	2021	96.6	80.5	95.9	58.9	97.2	94.1	87.4	82.9	92.4	35.2
RG-GCN [51]	2022	96.5	74.5	**98.2**	**79.4**	91.8	86.1	72.4	69.9	82.1	16.0
Stratified Transformer [41]	2022	96.7	81.2	95.9	57.9	97.3	93.7	87.6	84.0	93.9	39.3
NeiEA-Net [46]	2023	**97.0**	80.9	97.1	66.9	97.3	93.0	**97.3**	83.4	93.4	43.1
EyeNet [11]	2023	94.6	81.1	97.0	65.0	**97.8**	93.5	86.8	**84.9**	**94.0**	30.0
TCFAP-Net	2023	**97.0**	81.9	97.1	64.8	97.2	**94.3**	87.9	81.9	93.0	38.6

Table 5
Quantitative results of ablation studies on S3DIS using OA and mIoU metrics.

Variant	BFO	TFF	AP	S3DIS	
				OA	mIoU
Baseline				87.2	62.4
A1	✓			87.3	63.4
A2	✓	✓		88.1	64.2
Full Model	✓	✓	✓	**88.9**	**66.1**

Table 6
Ablation study on the sequences and combinations of self-attention and cross-attention in the TFF unit.

Variant	Setting	mIoU (%)
B1	self-attention	59.2
B2	cross-attention	62.8
B3	self-attention + self-attention	60.4
B4	cross-attention + cross-attention	61.6
B5	cross-attention + self-attention	63.6
Full Model	self-attention + cross-attention	**66.1**

showed an increase of approximately 1.0% on the A1. In summary, the entire encoder module exhibited a significant enhancement of approximately 2.0% compared to the baseline. The improvement in the AP module is also remarkable, with an increase of around 2.0%.

4.4.2. Ablation study of the attention mechanism

In this section, experiments are conducted on the sequences and combinations of the attention mechanism. The numerical results comparison between B1 and B2 from Table 6 indicates that cross-attention is superior to self-attention by approximately 3.0% when using a single attention mechanism. Similarly, comparing B3 and B4 leads to the same conclusion. This indicates that the mutual guidance of geometric and semantic information is advantageous for segmentation. The lower performance of B5 compared to the Full Model implies that self-attention is effective in filtering out anomalous features. Therefore, applying cross-attention after self-attention ensures correct guidance between geometry and semantics.

4.4.3. Ablation study of cross-feature interactions

Bilateral features (semantic and geometric features) can be regarded as two important properties for describing the raw point cloud, allowing more effective semantic segmentation. In this paper, we leverage the BFO unit and cross-attention to achieve the fusion of high-level semantic features and low-level geometric features for cross-feature interactions, improving segmentation accuracy. To verify the effectiveness of cross-feature interactions, we perform the following ablation experiments and record the results in Table 7. Comparing variants C1, C2, and C3 with our Full Model, our cross-feature interactions using the BFO unit and cross-attention can help our method yield

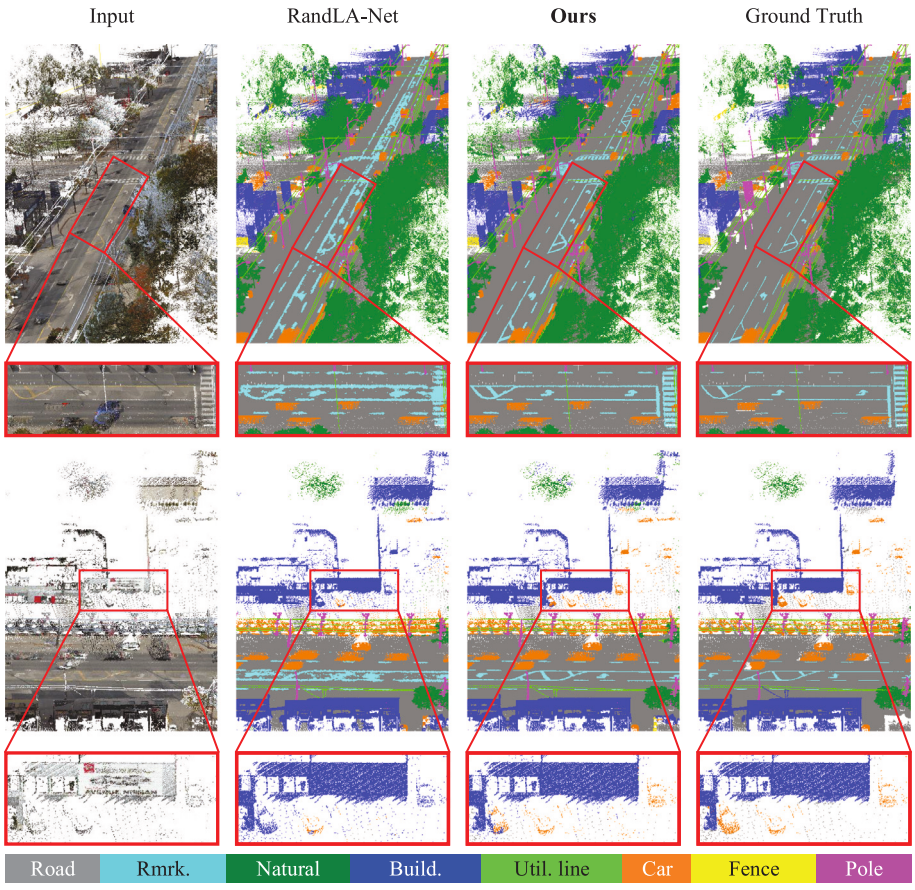

Fig. 9. Visual comparison of RandLA-Net (baseline) with our approach TCFAP-Net on Toronto3D.

Table 7
Ablation study for cross-feature interactions.

Variant	Setting	mIoU (%)
C1	removing BFO unit	65.2
C2	removing cross-attention in TFF unit	59.2
C3	removing BFO unit and cross-attention	59.0
Full Model	using BFO unit and cross-attention	**66.1**

Table 8
Effects of parameter K and L. When L is set to 3, experiments are not conducted due to hardware limitations.

Parameters	Settings	mIoU (%)	OA (%)	mAcc (%)
	8	64.2	87.0	70.6
	10	64.5	87.4	70.8
	12	65.0	88.2	71.2
K	14	65.8	89.0	71.7
	16	**66.1**	**88.9**	**72.8**
	18	65.7	87.4	72.1
	20	65.1	87.6	71.0
	0	64.2	88.1	70.6
L	1	65.8	88.2	71.0
	2	**66.1**	**88.9**	**72.8**

the best segmentation results. The above ablation studies confirm the effectiveness of our cross-feature interactions.

4.5. Discussions

Learning process charts. The learning curves of TCFAP-Net and RandLA-Net are illustrated in Fig. 10; our TCFAP-Net converges more quickly and exhibits more stability on both datasets compared to the baseline RandLA-Net. For instance, our network reaches convergence in the 75th epoch, but the baseline requires 150 epochs to converge, as seen by the S3DIS training loss curves.

Effects of parameters K and L. First, we conduct a thorough analysis of the setting for the K value in the KNN search. Through a comprehensive analysis of Table 8 and Fig. 11, we observe a notable upward trend in all three evaluation metrics as the K value increases from 8 to 16. Specifically, mIoU, OA, and mAcc show increases of 1.9%, 1.9%, and 2.2%, respectively. This is mainly due to the expansion of the receptive field of the central point as the K value increases, allowing for the acquisition of richer geometric and semantic features from neighboring points. However, once the K value exceeds a certain threshold, all three evaluation metrics exhibit a noticeable downward trend. This is because an excessively large receptive field may lead to the issue of information blurring when aggregating the information to the central point. Finally, we set the K value to 16 in our proposed method.

Then, a series of experiments are performed to investigate the effect caused by the number of changes of CFC units (denoted as L) in the AP module. Table 8 and Fig. 11 illustrate that as L increases, there is a noticeable improvement in mIoU, OA, and mAcc. In particular, mIoU shows a significant improvement, reaching 1.9%. This upward trend demonstrates the effectiveness of our module. However, it is important to note that the increase in L, is accompanied by a corresponding increase in computational resources. Therefore, after careful consideration, we ultimately set L to 2, striking a balance between maintaining effectiveness and controlling computational overhead.

Model complexity and efficiency. Model complexity and efficiency of the tested methods are critical to practical applications. Thus, we evaluate the model complexity (including the total number

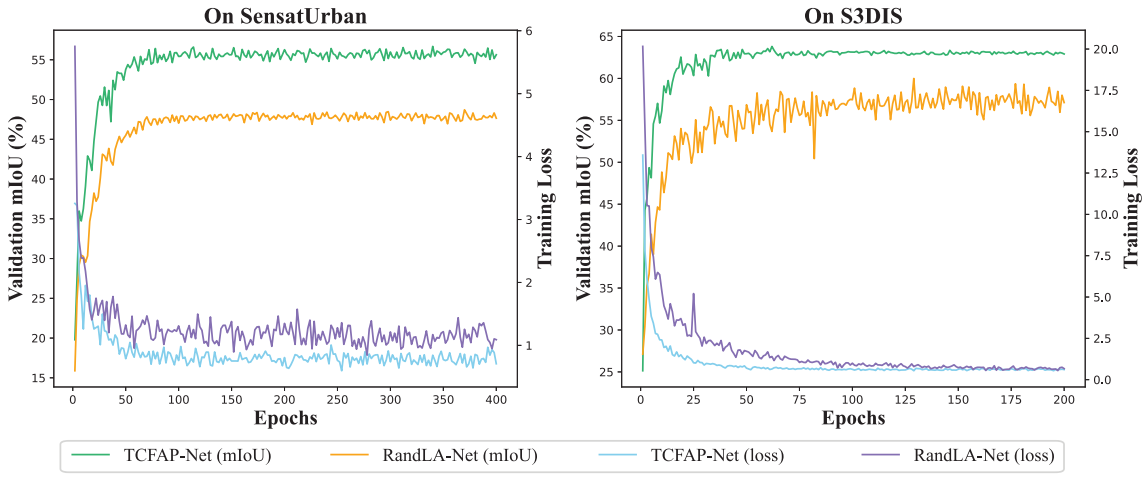

Fig. 10. Validation mIoU and training loss curves of our TCFAP-Net and RandLA-Net on S3DIS datasets.

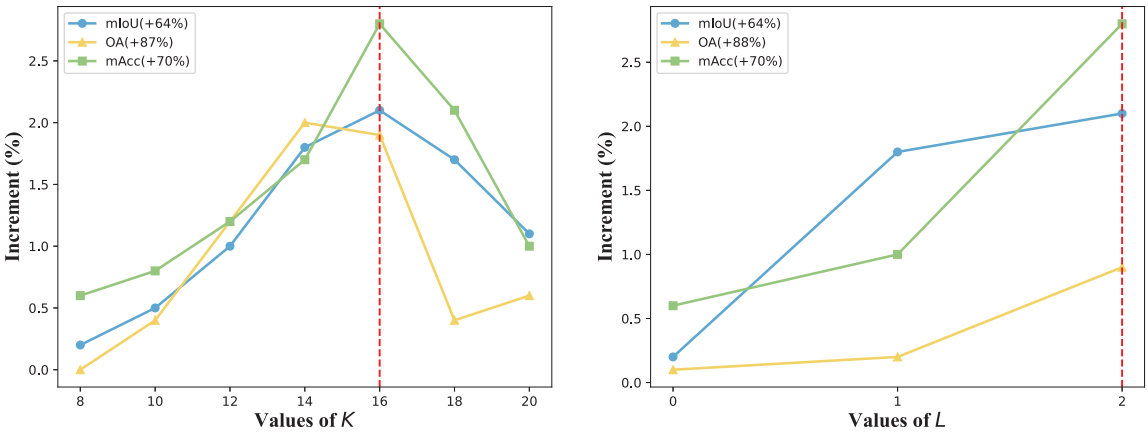

Fig. 11. Visual curves of mIoU, OA, and mAcc for different settings of K values in KNN and the number of CFC units L.

of parameters abbreviated as Params. and floating point operation per second abbreviated as FLOPs) and efficiency (including the per-batch training time and inference time) of the tested methods and record the results in Table 9. The same set of parameters is utilized in both the training and testing stages for fairness. As we can see, RandLA-Net has the fewest parameters and the highest efficiency among all methods, but its segmentation accuracy is the lowest. BAF-LAC improves segmentation accuracy somewhat over RandLA-Net, but its complexity and efficiency increase accordingly. Although Point Transformer and Stratified Transformer provide great segmentation accuracy on S3DIS (6-Fold), they come with larger parameters and longer inference time. Compared to these methods, our approach strikes a balance between model effectiveness, complexity, and efficiency. For model complexity, since our approach is based on the Transformer architecture, it has the same high parameters as Point Transformer and Stratified Transformer. However, the FLOPs of our network are lower than those of Point Transformer and Stratified Transformer due to factors such as the number of Transformer blocks, the feature dimensions of the encoder and decoder, and the sampling strategy. For model efficiency, our approach's per-batch and inference time is approximately 1/6 of Point Transformer and Stratified Transformer. This is because random sampling employed in our approach is more efficient in processing large-scale point clouds than farthest point sampling employed in Point Transformer and Stratified Transformer. Due to the high complexity of Point Transformer and Stratified Transformer, they are unsuitable for dealing with large-scale point clouds, such as Sensaturban, which contains 10^9 points. In contrast, S3DIS and Toronto3D only include 10^8 and 10^7 points, respectively. Simultaneously, our approach performs well in terms of effectiveness, ranking first and second in segmentation accuracy on SensatUrban and Toronto3D, respectively, while behind just marginally on S3DIS (6-Fold). Considering comprehensively the performance of the tested methods in terms of effectiveness, complexity, and efficiency, our approach achieves impressive results.

Effects of different sampling strategies. We evaluate the performance of our framework using different sampling strategies and record the results in Table 10. As we can see, our method outperforms Point Transformer in segmentation performance and inference time using the same sampling strategy. Although employing FPS can improve segmentation performance over RS, it requires more computational resources. Therefore, we adopt RS as our sampling strategy to balance model effectiveness and efficiency.

Limitation. We discuss the limitations of our method in two aspects. First, the architecture relies on KNN search to learn local semantic contexts. Since the point distribution in point clouds may vary significantly between places, a fixed number of nearest neighbors cannot provide sufficient contextual information, especially in complex places. Second, although the random sampling strategy is efficient, it may fail to learn structural characteristics, resulting in unsatisfactory results for structural indoor environments. Therefore, promising future directions include exploring flexible neighbor searching and more effective sampling strategies.

5. Conclusion

In this paper, we introduce a scene-level semantic segmentation approach of point clouds based on Transformer. Our proposed approach,

Table 9
Model complexity and efficiency analysis of TCFAP-Net and the compared methods.

Method	Params. (M)	FLOPs (G)	Per-batch time (ms)	Inference time (s/(10^6 points))	mIoU (%) SensatUrban	Tornoto3D	S3DIS (6-Fold)
RandLA-Net [9]	**4.99**	5.80	**183.27**	**73.13**	52.7	81.8	70.0
BAF-LAC [30]	6.39	6.62	246.94	93.10	54.1	**82.0**	71.7
Point Transformer [39]	7.77	5.64	1385.62	513.49	–	80.5	73.5
Stratified Transformer [41]	8.02	6.35	1747.06	647.43	–	81.2	**73.7**
TCFAP-Net	8.10	**5.25**	280.65	100.45	**64.1**	81.9	72.5

Table 10
Quantitative results of our method and Point Transformer with different sampling strategies on S3DIS (6-fold).

Sampling strategy	Method	mIoU (%)	OA (%)	Inference time
RS	Point Transformer	71.8	88.9	108.57
	TCFAP-Net	72.5	89.3	100.45
FPS	Point Transformer	73.5	90.2	513.49
	TCFAP-Net	73.5	90.4	496.08

RS: Random Sampling; FPS: Farthest Point Sampling.

coined as TCFAP-Net, comprises two core modules: Transformer-based Cross-feature Fusion (TCF) and Adaptive Perception (AP). On the one hand, TCF can learn discriminative fusion features to resolve the boundary mis-segmentation and blurring problem. On the other hand, AP can adaptively perceive long-range contextual information from varying feature scales. The comprehensive experiments demonstrate the superiority of our approach quantitatively and qualitatively. In particular, compared with the competing segmentation methods, our approach achieves state-of-the-art performance on the tested benchmark datasets (SensatUrban, S3DIS, and Toronto3D). In future work, we would like to further explore the potential of our approach and extend its application to even more datasets for diverse 3D tasks. In addition, we will focus on researching methods to enhance the model's generalization, such as utilizing domain adaptation techniques to transfer pre-trained models from the source domain to the target domain in an unsupervised or semi-supervised manner.

CRediT authorship contribution statement

Jianjun Zhang: Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Zhipeng Jiang:** Writing – original draft, Validation, Methodology, Data curation. **Qinjun Qiu:** Writing – review & editing, Validation, Data curation. **Zheng Liu:** Writing – review & editing, Project administration, Methodology.

Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this manuscript and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

Data availability

We have shared the code link in the abstract part of the paper.

Acknowledgments

This work was supported by National Key R&D Program of China (2022YFB3904100), Guangdong Enterprise Key Laboratory for Urban Sensing, Monitoring and Early Warning, China (2020B121202019).

References

[1] Z. Liu, Y. Zhao, S. Zhan, Y. Liu, R. Chen, Y. He, PCDNF: Revisiting learning-based point cloud denoising via joint normal filtering, IEEE Trans. Vis. Comput. Graphics (2023) http://dx.doi.org/10.1109/TVCG.2023.3292464.

[2] Z. Liu, X. Xin, Z. Xu, W. Zhou, C. Wang, R. Chen, Y. He, Robust and accurate feature detection on point clouds, Comput. Aided Des. 164 (2023) http://dx.doi.org/10.1016/j.cad.2023.103592.

[3] X. Wen, Z. Han, G. Youk, Y.-S. Liu, CF-SIS: Semantic-instance segmentation of 3D point clouds by context fusion with self-attention, in: Proceedings of the ACM International Conference on Multimedia, ACM MM, 2020, pp. 1661–1669, http://dx.doi.org/10.1145/3394171.3413829.

[4] G. Yang, F. Xue, Q. Zhang, K. Xie, C.-W. Fu, H. Huang, UrbanBIS: a large-scale benchmark for fine-grained urban building instance segmentation, in: ACM SIGGRAPH Conference Proceedings, 2023, pp. 1–11, http://dx.doi.org/10.1145/3588432.3591508.

[5] T. Zhao, L. Li, T. Tian, J. Ma, J. Tian, Patch-guided point matching for point cloud registration with low overlap, Pattern Recognit. 144 (2023) 109876, http://dx.doi.org/10.1016/j.patcog.2023.109876.

[6] K. Slimani, C. Achard, B. Tamadazte, RoCNet++: Triangle-based descriptor for accurate and robust point cloud registration, Pattern Recognit. 147 (2024) 110108, http://dx.doi.org/10.1016/j.patcog.2023.110108.

[7] Y. Li, Z. Zhao, J. Fan, W. Li, ADR-MVSNet: A cascade network for 3D point cloud reconstruction with pixel occlusion, Pattern Recognit. 125 (2022) 108516, http://dx.doi.org/10.1016/j.patcog.2021.108516.

[8] B. Ma, Y.-S. Liu, M. Zwicker, Z. Han, Surface reconstruction from point clouds by learning predictive context priors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 6326–6337.

[9] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, A. Markham, RandLA-Net: Efficient semantic segmentation of large-scale point clouds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 11105–11114, http://dx.doi.org/10.1109/CVPR42600.2020.01112.

[10] M. Li, Y. Xie, L. Ma, Paying attention for adjacent areas: Learning discriminative features for large-scale 3D scene segmentation, Pattern Recognit. 129 (2022) 108722, http://dx.doi.org/10.1016/j.patcog.2022.108722.

[11] S. Yoo, Y. Jeong, M. Jameela, G. Sohn, Human vision based 3D point cloud semantic segmentation of large-scale outdoor scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 6577–6586, http://dx.doi.org/10.1109/CVPRW59228.2023.00699.

[12] R.Q. Charles, H. Su, M. Kaichun, L.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, CVPR, 2017, pp. 77–85, http://dx.doi.org/10.1109/CVPR.2017.16.

[13] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI, 2015, pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[14] S. Fan, Q. Dong, F. Zhu, Y. Lv, P. Ye, F.-Y. Wang, SCF-Net: Learning spatial contextual features for large-scale point cloud segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 14499–14508, http://dx.doi.org/10.1109/CVPR46437.2021.01427.

[15] S. Qiu, S. Anwar, N. Barnes, Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 1757–1767, http://dx.doi.org/10.1109/CVPR46437.2021.00180.

[16] X. Yang, D. Zhou, S. Liu, J. Ye, X. Wang, Deep model reassembly, in: Proceedings of the Annual Conference on Neural Information Processing Systems, Vol. 35, NIPS, 2022, pp. 25739–25753.
[17] X.W. Xingyi Yang, Factorizing knowledge in neural networks, in: Proceedings of the European Conference on Computer Vision, ECCV, 2022.
[18] S. Liu, J. Ye, R. Yu, X. Wang, Slimmable dataset condensation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 3759–3768, http://dx.doi.org/10.1109/CVPR52729.2023.00366.
[19] S. Liu, K. Wang, X. Yang, J. Ye, X. Wang, Dataset distillation via factorization, in: Proceedings of the Annual Conference on Neural Information Processing Systems, Vol. 35, NIPS, 2022, pp. 1100–1113.
[20] X. Yang, D. Zhou, J. Feng, X. Wang, Diffusion probabilistic model made slim, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 22552–22562, http://dx.doi.org/10.1109/CVPR52729.2023.02160.
[21] Z. Wang, F. Lu, VoxSegNet: Volumetric CNNs for semantic part segmentation of 3D shapes, IEEE Trans. Vis. Comput. Graphics 26 (2020) 2919–2930, http://dx.doi.org/10.1109/TVCG.2019.2896310.
[22] M. Liu, Q. Zhou, H. Zhao, J. Li, Y. Du, K. Keutzer, L. Du, S. Zhang, Prototype-voxel contrastive learning for LiDAR point cloud panoptic segmentation, in: IEEE International Conference on Robotics and Automation, ICRA, 2022, pp. 9243–9250, http://dx.doi.org/10.1109/ICRA46639.2022.9811638.
[23] T. Yu, J. Meng, J. Yuan, Multi-view harmonized bilinear network for 3D object recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 186–194, http://dx.doi.org/10.1109/CVPR.2018.00027.
[24] A.H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, O. Beijbom, PointPillars: Fast encoders for object detection from point clouds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 12689–12697, http://dx.doi.org/10.1109/CVPR.2019.01298.
[25] C.R. Qi, L. Yi, H. Su, L.J. Guibas, PointNet++: Deep hierarchical feature learning on point sets in a metric space, in: Proceedings of the Annual Conference on Neural Information Processing Systems, NIPS, 30, 2017, pp. 5105–5114, https://dl.acm.org/doi/10.5555/3295222.3295263.
[26] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, B. Chen, PointCNN: Convolution on X-Transformed points, in: Proceedings of the Annual Conference on Neural Information Processing Systems, vol. 31, NIPS, 2018, pp. 828–838, https://dl.acm.org/doi/10.5555/3326943.3327020.
[27] H. Thomas, C.R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, L. Guibas, KPConv: Flexible and deformable convolution for point clouds, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2019, pp. 6410–6419, http://dx.doi.org/10.1109/ICCV.2019.00651.
[28] Y. Jing, Y. Yang, X. Wang, M. Song, D. Tao, Amalgamating knowledge from heterogeneous graph neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 15704–15713, http://dx.doi.org/10.1109/CVPR46437.2021.01545.
[29] Y. Jing, C. Yuan, L. Ju, Y. Yang, X. Wang, D. Tao, Deep graph reprogramming, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 24345–24354, http://dx.doi.org/10.1109/CVPR52729.2023.02332.
[30] H. Shuai, X. Xu, Q. Liu, Backward attentive fusing network with local aggregation classifier for 3D point cloud semantic segmentation, IEEE Trans. Image Process. 30 (2021) 4973–4984, http://dx.doi.org/10.1109/TIP.2021.3073660.
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the Annual Conference on Neural Information Processing Systems, NIPS, 2017, pp. 6000–6010, https://dl.acm.org/doi/10.5555/3295222.3295349.
[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for image recognition at scale, in: International Conference on Learning Representations, ICLR, 2021.
[33] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2021, pp. 10012–10022, http://dx.doi.org/10.1109/ICCV48922.2021.00986.
[34] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, C. Shen, Twins: Revisiting the design of spatial attention in vision transformers, in: Proceedings of the Annual Conference on Neural Information Processing Systems, NIPS, 2021.
[35] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, B. Guo, CSWin Transformer: A general vision transformer backbone with cross-shaped windows, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 12124–12134, http://dx.doi.org/10.1109/CVPR52688.2022.01181.
[36] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, J. Gao, Focal attention for long-range interactions in vision transformers, in: Proceedings of the Annual Conference on Neural Information Processing Systems, Vol. 34, NIPS, 2021, pp. 30008–30022.
[37] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, L. Soler, U-Net Transformer: Self and cross attention for medical image segmentation, in: Machine Learning in Medical Imaging, MLMI, 2021, pp. 267–276.
[38] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R.R. Martin, S.-M. Hu, PCT: Point cloud transformer, Comp. Visual Media 7 (2021) 187–199, http://dx.doi.org/10.1007/s41095-021-0229-5.
[39] H. Zhao, L. Jiang, J. Jia, P. Torr, V. Koltun, Point Transformer, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2021, pp. 16239–16248, http://dx.doi.org/10.1109/ICCV48922.2021.01595.
[40] X. Wu, Y. Lao, L. Jiang, X. Liu, H. Zhao, Point Transformer V2: Grouped vector attention and partition-based pooling, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Proceedings of the Annual Conference on Neural Information Processing Systems, Vol. 35, NIPS, 2022, pp. 33330–33342.
[41] X. Lai, J. Liu, L. Jiang, L. Wang, H. Zhao, S. Liu, X. Qi, J. Jia, Stratified transformer for 3D point cloud segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 8490–8499, http://dx.doi.org/10.1109/CVPR52688.2022.00831.
[42] Q. Hu, B. Yang, S. Khalid, W. Xiao, N. Trigoni, A. Markham, Towards Semantic Segmentation of Urban-Scale 3D Point Clouds: A dataset, benchmarks and challenges, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 4975–4985, http://dx.doi.org/10.1109/CVPR46437.2021.00494.
[43] M. Tatarchenko, J. Park, V. Koltun, Q.-Y. Zhou, Tangent convolutions for dense prediction in 3D, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 3887–3896, http://dx.doi.org/10.1109/CVPR.2018.00409.
[44] L. Landrieu, M. Simonovsky, Large-scale point cloud semantic segmentation with superpoint graphs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 4558–4567, http://dx.doi.org/10.1109/CVPR.2018.00479.
[45] B. Graham, M. Engelcke, L.v.d. Maaten, 3D semantic segmentation with submanifold sparse convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 9224–9232, http://dx.doi.org/10.1109/CVPR.2018.00961.
[46] Y. Xu, W. Tang, Z. Zeng, W. Wu, J. Wan, H. Guo, Z. Xie, NeiEA-NET: Semantic segmentation of large-scale point cloud scene via neighbor enhancement and aggregation, Int. J. Appl. Earth Obs. Geoinf. 119 (2023) 103285, http://dx.doi.org/10.1016/j.jag.2023.103285.
[47] I. Armeni, O. Sener, A.R. Zamir, H. Jiang, I. Brilakis, M. Fischer, S. Savarese, 3D semantic parsing of large-scale indoor spaces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 1534–1543, http://dx.doi.org/10.1109/CVPR.2016.170.
[48] W. Tan, N. Qin, L. Ma, Y. Li, J. Du, G. Cai, K. Yang, J. Li, Toronto-3D: A large-scale mobile LiDAR dataset for semantic segmentation of urban roadways, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 797–806, http://dx.doi.org/10.1109/CVPRW50498.2020.00109.
[49] Z. Du, H. Ye, F. Cao, A novel local-global graph convolutional method for point cloud semantic segmentation, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–15, http://dx.doi.org/10.1109/TNNLS.2022.3155282.
[50] J. Du, G. Cai, Z. Wang, S. Huang, J. Su, J. Marcato Junior, J. Smit, J. Li, ResDLPS-Net: Joint residual-dense optimization for large-scale point cloud semantic segmentation, ISPRS J. Photogramm. Remote Sens. 182 (2021) 37–51, http://dx.doi.org/10.1016/j.isprsjprs.2021.09.024.
[51] Z. Zeng, Y. Xu, Z. Xie, J. Wan, W. Wu, W. Dai, RG-GCN: A random graph based on graph convolution network for point cloud semantic segmentation, Remote Sens. 14 (2022b) http://dx.doi.org/10.3390/rs14164055.

Jianjun Zhang received the B.S. degree in computer science and technology from Beijing Institute of Petrochemical Technology, Beijing, China, in 2022 and is currently a M.S. candidate in China University of Geosciences, Wuhan, China. His research interests include 3D vision, 3D deep learning and pattern recognition.

Zhipeng Jiang received the B.S. degree in computer science and technology from Beijing Institute of Petrochemical Technology, Beijing, China, in 2022 and is currently a M.S. candidate at China University of Geosciences, Wuhan, China. His research interests include 3D vision, 3D deep learning and computer graphics.

Qinjun Qiu received the B.S. degree and the M.S. degree from the China Three Gorges University, Yichang, China, in 2011 and 2014, respectively, and the Ph.D. degree from the China University of Geosciences, Wuhan, China, in 2020. His research interests include deep learning, text mining, and knowledge graph.

Zheng Liu is currently an associate professor at China University of Geosciences (Wuhan). He received the Ph.D. degree from Central China Normal University in 2012. From 2013 to 2014, he held a post-doctoral position at University of Science and Technology of China. His research interests include geometry processing, 3D vision, 3D deep learning, and computer graphics.