Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.078692

ARTICLE

# PointNMSA: An Improved PointNeXt Network with Non-Local Multi-Scale Aggregation for 3D Point Cloud Semantic Segmentation

**Aihua Wu and Chenlu Huang***

College of Information Engineering, Shanghai Maritime University, Shanghai, China

*Corresponding Author: Chenlu Huang. Email: 202430310333@stu.shmtu.edu.cn

Received: 06 January 2026; Accepted: 20 April 2026; Published: 15 June 2026

**ABSTRACT:** Three-dimensional (3D) point cloud semantic segmentation is a core task in indoor scene understanding, providing detailed semantic information about spatial structures and object categories in indoor environments. Although methods based on deep learning have made steady progress in recent years, accurately segmenting complex indoor scenes remains challenging due to the unordered nature of point clouds and variations across large scales. Most existing networks have limited capability for multi-scale feature aggregation and struggle to balance local geometric details with global semantic context. These issues are further exacerbated by hierarchical downsampling, which often leads to the loss of fine-grained structural information. Moreover, feature interaction restricted to local neighborhoods may limit the capture of non-local semantic dependencies in complex indoor scenes. To address these limitations, we propose PointNMSA (PointNeXt with Non-local Multi-Scale Aggregation), an improved semantic segmentation network built upon the PointNeXt backbone. A Multi-Scale Feature Enhancement (MSFE) module is introduced in the decoding stage to fuse features from different encoding levels, and further refines the fused features to produce more stable multi-scale representations, which preserves geometric details across scales. In addition, a Convolution-Attention Mixing (CA-Mix) module is designed to jointly integrate local spatial structures and non-local contextual dependencies via dual-stream aggregation and multi-dimensional attention fusion, thereby enabling more discriminative feature representations. Experiments on the Stanford Large-Scale 3D Indoor Spaces (S3DIS) benchmark demonstrate the effectiveness of PointNMSA. On the Area 5 test split, PointNMSA achieves a mean intersection over union (mIoU) of 65.10%, outperforming the PointNeXt baseline by 1.59%, while introducing only a modest increase in computational cost (latency from 42.24 to 45.18 ms and parameters from 3.16 to 8.67M). Despite the noticeable growth in parameter count, the increase in inference latency remains relatively limited, indicating a favorable trade-off between segmentation accuracy and computational efficiency. Additional cross-dataset experiments on ScanNet further verify that PointNMSA maintains stable gains under different indoor scene distributions. Such performance gains suggest that PointNMSA provides a more robust and generalizable solution for semantic segmentation in large-scale indoor environments with complex structural layouts.

**KEYWORDS:** 3D point cloud semantic segmentation; indoor scene understanding; multi-scale feature aggregation; non-local context integration; PointNeXt

## 1 Introduction

Semantic segmentation of 3D point clouds is an essential task in indoor scene understanding. By assigning semantic labels to individual points, it provides detailed information about spatial structures and object categories, supporting scene interpretation and environment interaction in indoor spaces. The widespread use of Red-Green-Blue-Depth (RGB-D) sensors and Light Detection and Ranging (LiDAR)

sensors has led to the availability of large-scale indoor point cloud datasets. However, accurate segmentation remains challenging due to the unordered nature of point sets, irregular spatial distributions, and variations in point density. Indoor environments often contain complex structural layouts, closely arranged objects, and categories with similar geometric characteristics, which make reliable classification difficult.

Early approaches mainly relied on hand-crafted geometric features and traditional algorithms such as clustering, region growing, and Random Sample Consensus (RANSAC) [1]. These methods can achieve reasonable performance in relatively regular scenes, but they are highly sensitive to noise, occlusion, and complex geometric structures, which limit their generalization ability. With the development of deep learning, the research paradigm has gradually shifted toward end-to-end data-driven methods. Several mainstream approaches have been established, including point-based methods represented by PointNet [2] and PointNet++ [3], voxel-based methods such as VoxelNet [4] and MinkowskiNet [5], as well as graph-based methods represented by Dynamic Graph Convolutional Neural Network (DGCNN) [6]. These approaches alleviate the unordered nature of point clouds to some extent and have improved semantic segmentation performance in recent years [7]. However, accurate recognition of small objects, thin structures, and object boundaries is still challenging. In addition, effectively combining fine-grained local geometry with broader spatial context remains difficult. These challenges motivate more effective multi-scale feature aggregation and non-local context modeling strategies for indoor scenes.

As illustrated in Fig. 1, PointNeXt [8] revisits the PointNet++ backbone with inverted residual multi-layer perceptron (MLP) blocks and refined training strategies, achieving competitive segmentation performance while maintaining high computational efficiency. And it has become a strong baseline for semantic segmentation. However, similar to earlier point-based architectures [2,3], its feature learning still relies primarily on local neighborhood aggregation. Although hierarchical downsampling enlarges the receptive field, feature interaction at each stage remains restricted to predefined local regions. In complex indoor scenes with non-uniform density distributions and multi-scale structural variations [9–11], relying primarily on local neighborhood aggregation may limit the coordination between fine geometric details and higher-level semantic information, particularly in thin structures and boundary regions where consistent contextual cues are essential. Recent transformer-based approaches attempt to address this issue by enabling feature interaction beyond fixed local neighborhoods, allowing semantically related points to contribute to each other even when they are not spatially adjacent. While this broader interaction improves contextual modeling, fully global designs often introduce additional computational cost and may weaken the efficiency advantages of lightweight backbones. Therefore, it is necessary to simultaneously enhance multi-scale feature coordination and non-local contextual interaction within an efficiency-oriented architecture, rather than relying solely on purely local aggregation or fully global attention mechanisms. Based on these considerations, we propose PointNMSA, a non-local multi-scale aggregation framework built upon the PointNeXt backbone, which introduces structured non-local interaction while preserving the efficiency characteristics of the original architecture. It incorporates feature interaction beyond local neighborhoods at each hierarchical stage and integrates cross-level feature enhancement during feature propagation, thereby improving semantic consistency across different resolutions while maintaining computational efficiency. The main contributions of this work are summarized as follows:

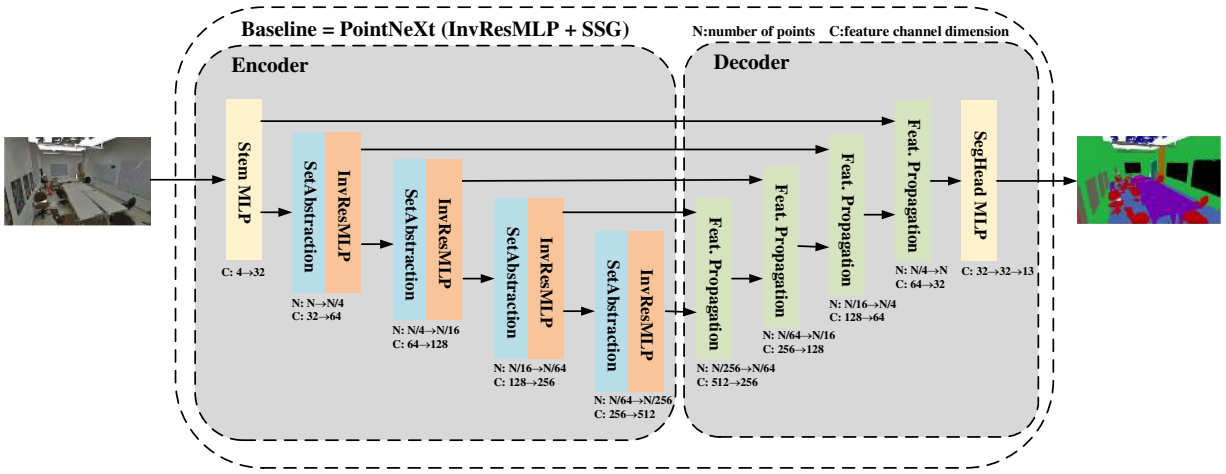


**Figure 1:** PointNeXt network architecture. Adapted with permission from reference [8].

(1) A Multi-Scale Feature Enhancement (MSFE) module is introduced in the decoding stage for cross-level feature fusion and feature refinement. Instead of directly concatenating features from different layers, MSFE adaptively integrates encoded features at different scales to alleviate the loss of fine-grained geometric information caused by downsampling, which improves boundary delineation and the representation of sparse structures. After fusion, MSFE further enhances the aggregated features with a refinement unit, producing more stable and informative features for subsequent propagation.

(2) A Convolution-Attention Mixing (CA-Mix) module is introduced to enhance feature interaction beyond local neighborhoods while preserving local geometric structures. Rather than using global self-attention alone, CA-Mix combines convolutional aggregation with contextual attention to coordinate local details and broader semantic information. This design alleviates the limitations of purely local feature learning and improves prediction consistency in complex indoor scenes.

(3) Extensive experiments are conducted on the S3DIS dataset [9] to evaluate the effectiveness of PointNMSA. The ablation studies analyze the contributions of the multi-scale feature enhancement and non-local context integration components, as well as the internal design of their sub-modules, demonstrating their complementary roles in improving segmentation performance. Furthermore, comparative experiments and per-class analyses validate the effectiveness of the proposed method in both overall performance and category-level segmentation accuracy. In addition, cross-dataset experiments on ScanNet [10] further verify the generalization capability of PointNMSA under different indoor scene distributions.

## 2 Related Work

In recent years, point cloud semantic segmentation has attracted extensive research attention. Existing studies mainly focus on key aspects such as local geometric feature representation, multi-scale feature aggregation, and global context integration. Among these directions, multi-scale feature aggregation and non-local context integration are widely considered effective for improving segmentation performance in complex indoor scenes. Related research along these lines has continued to progress in recent years.

### *2.1 Multi-Scale Feature Aggregation*

Point clouds exhibit non-uniform spatial distributions and varying scale characteristics. Under such conditions, single-scale feature representations are insufficient to capture both local geometric details and

high-level semantic information. As a result, multi-scale feature aggregation has become essential. Existing methods generally follow two strategies for multi-scale processing. One category of methods introduces multi-scale receptive fields within local neighborhoods. This is commonly achieved through mechanisms such as multi-scale grouping or parallel feature extraction to enhance local geometric feature representations, as explored in PointNet++ [3], KPConv [12] and PointConv [13]. Recent studies have further explored adaptive multi-scale feature aggregation by leveraging voxel-based context and point-level representations, enabling more flexible receptive field adaptation for point cloud semantic segmentation [14,15]. Another category focuses on cross-level feature fusion within encoder–decoder architectures, where features at different resolutions are integrated through skip connections or feature concatenation. Representative models following this strategy include RandLA-Net [16], PointNeXt [8] and U-Next [17].

Although the above methods alleviate the loss of fine-grained information caused by downsampling to some extent, most of them rely on fixed feature fusion strategies. Such designs make it difficult to balance feature contributions across scales under varying scene structures. Moreover, some multi-scale approaches depend on parallel branches or introduce additional parameters, leading to increased model complexity. Although segmentation accuracy is improved, this increase in complexity limits their suitability for direct integration into efficiency-oriented baseline networks. A key challenge, therefore, lies in enabling more flexible multi-scale feature aggregation while maintaining model efficiency.

### *2.2 Attention Mechanisms and Non-Local Feature Integration*

To enhance the capture of long-range dependencies and global contextual information, attention mechanisms have been increasingly applied to point cloud semantic segmentation. Self-attention–based models explicitly capture global relationships among points and enable more flexible feature interactions. Representative approaches following this direction include Point Transformer [18], Point Cloud Transformer (PCT) [19], Stratified Transformer [20], and Swin3D [21,22], which have demonstrated competitive performance in semantic segmentation tasks.

However, global context integration based purely on self-attention often incurs high computational and memory costs [23]. This issue becomes more pronounced when processing large scale point cloud data, which limits scalability to some extent. Motivated by this issue, recent studies have explored more efficient and scalable attention designs to reduce the computational overhead of global context integration in point cloud semantic segmentation [24,25]. To further alleviate this problem, recent studies have explored combining attention mechanisms with local operators such as convolutions or MLPs. These approaches introduce non-local contextual information while preserving local spatial structures, aiming to balance performance and efficiency. Representative examples include Superpoint Transformer [26], which introduces an efficient superpoint representation to model contextual relationships in large scale point clouds, and Cross-Fusion Self-Attention Network (CFSA-Net) [27], which employs a cross-fusion self-attention mechanism to jointly capture local structures and long-range contextual dependencies. Although such hybrid methods reduce computational complexity, they still struggle to effectively coordinate local geometric features with global semantic information.

## 3 PointNMSA

Building upon PointNeXt [8], we construct an enhanced network termed PointNMSA, as illustrated in Fig. 2. PointNMSA is designed to address the limitations of PointNeXt in multi-scale feature aggregation and in handling non-uniform point distributions. To this end, two core components are integrated into the original framework: a Multi-Scale Feature Enhancement (MSFE) module and a Convolution-Attention Mixing (CA-Mix) module. MSFE performs cross-level multi-scale feature aggregation, while CA-Mix introduces

non-local contextual interaction within each abstraction stage. Their coordinated design forms a unified non-local multi-scale aggregation mechanism, enabling the network to enhance structural representation across scales and strengthen long-range semantic dependencies.

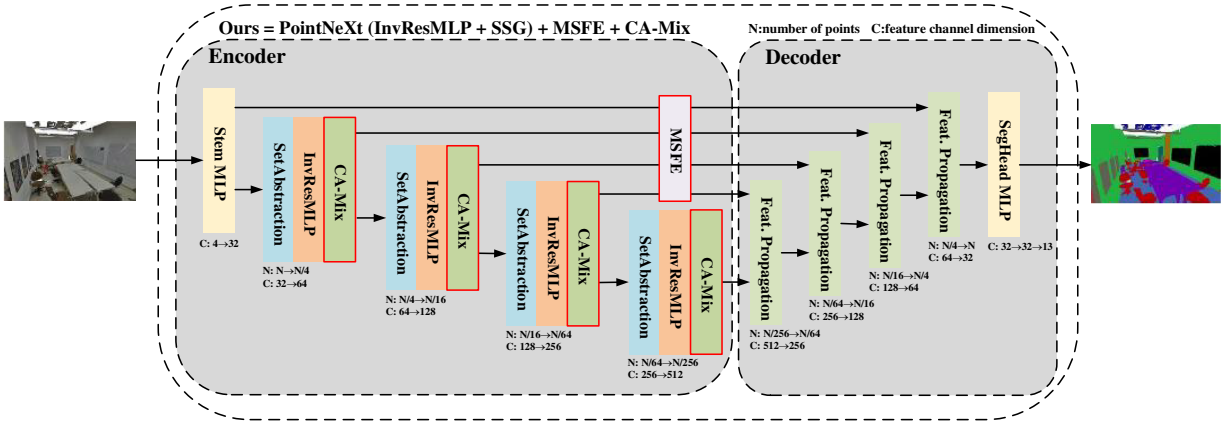


**Figure 2:** PointNMSA network architecture.

Fig. 2 shows the overall network pipeline of PointNMSA. The raw point cloud is first mapped to an initial feature space through an MLP. In the encoder, an Inverted Residual MLP (InvResMLP) and a CA-Mix module are sequentially stacked at each Set Abstraction stage. This design allows local geometric features to be extracted while facilitating non-local context integration. Features from different encoding levels are then fused by the MSFE module. Through cross-level fusion and adaptive weight assignment, the MSFE module integrates multi-scale contextual information and produces features that are globally informed and semantically consistent. Finally, multi-stage feature propagation is applied to progressively restore spatial resolution and perform point-wise prediction, yielding the final semantic segmentation results. With this architecture, PointNMSA preserves the efficiency of the PointNeXt framework while strengthening how the network represents geometry across scales and resolves semantic ambiguity in complex scenes.

### *3.1 Multi-Scale Feature Enhancement Module (MSFE)*

Although PointNeXt achieves a good balance between efficiency and performance, its single-scale grouping strategy in the encoder inevitably leads to the loss of fine-grained geometric information. This design also limits effective feature fusion across different receptive fields. These limitations become more apparent in indoor scenes with non-uniform point distributions or complex structures, where segmentation accuracy around object boundaries and thin structures is often degraded.

To alleviate these issues, we introduce a Multi-Scale Feature Enhancement (MSFE) module to replace and enhance the conventional cross-level feature propagation mechanism used in PointNeXt. As illustrated in Fig. 3, the MSFE module incorporates multi-scale feature aggregation and feature refinement in the decoding stage. By more effectively leveraging features from different encoding levels, the module supports the recovery of geometric details lost during downsampling and produces more informative representations for semantic segmentation.

The MSFE module consists of two cascaded sub-units: an Adaptive Multi-Scale Aggregation Block (AMSAB) and an Activation-Free Feature Refinement Block (AFFRB). The AMSAB adaptively fuses features from different scales, while the AFFRB further refines the fused features to achieve stable feature representations.

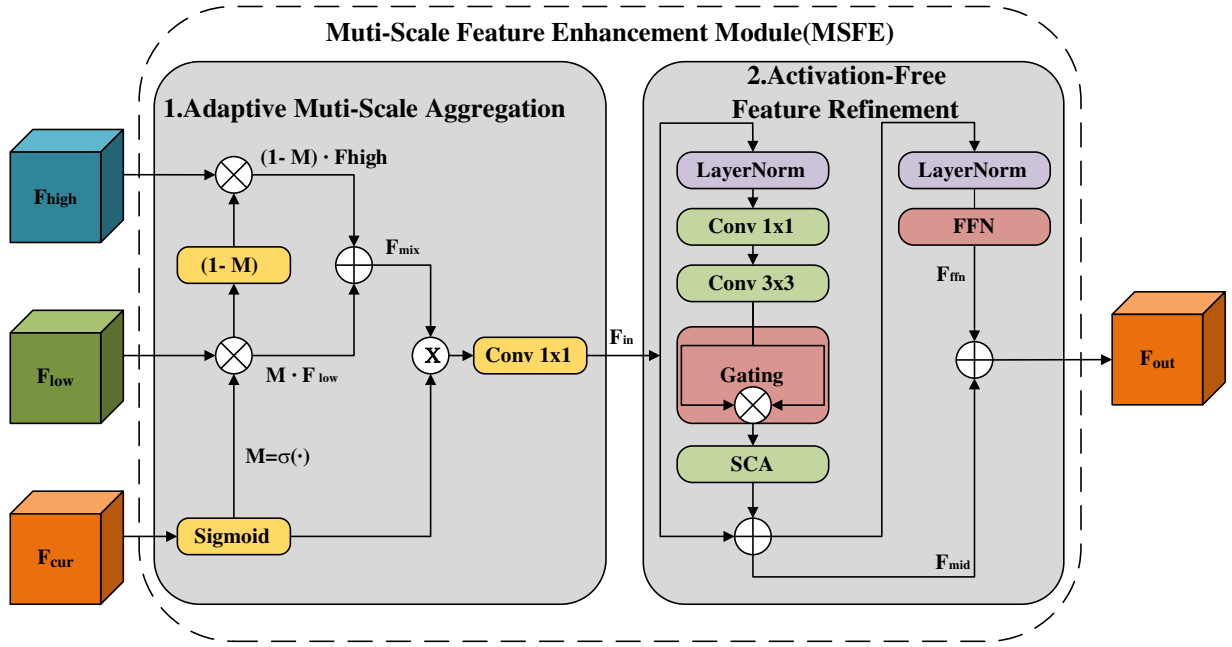


**Figure 3:** Architecture of the proposed Multi-Scale Feature Enhancement (MSFE) module.

*3.1.1 Adaptive Multi-Scale Aggregation Block (AMSAB)*

As illustrated on the left side of Fig. 3, the AMSAB receives three input feature streams: the current-level feature $F_{cur}$, a higher-level feature $F_{high}$ obtained through downsampling, and a lower-level feature $F_{low}$ obtained before downsampling. These three features are aligned in spatial resolution and channel dimension, allowing feature fusion at the same scale. To balance the contributions of features from different scales according to the structural information of the current level, the AMSAB first uses $F_{cur}$ as guidance to generate a spatially selective weighting mask $M$:

$$M = \sigma\left(Conv(F_{cur})\right) \tag{1}$$

Subsequently, the weighting mask $M$ and its complementary form are applied to perform complementary weighted fusion of $F_{low}$ and $F_{high}$, producing the fused feature $F_{mix}$:

$$F_{mix} = M \otimes F_{low} + (1 - M) \otimes F_{high} \tag{2}$$

Through this fusion strategy, high-frequency geometric details from the lower-level feature are emphasized in regions with larger weights, while low-frequency semantic information from the higher-level feature is introduced in the complementary regions. In this manner, multi-scale information is effectively combined in the spatial domain. After weighted fusion, the fused feature $F_{mix}$ is concatenated with the current-level feature $F_{cur}$ along the channel dimension and passed through a convolution operation to perform feature integration and channel compression, yielding the output feature $F_{in}$ of the AMSAB:

$$F_{in} = Conv(Concat(F_{mix}, F_{cur})) \tag{3}$$

With this design, the AMSAB enables cross-level multi-scale feature aggregation without introducing significant additional computational overhead.

*3.1.2 Activation-Free Feature Refinement Block (AFFRB)*

As illustrated on the right side of Fig. 3, the AFFRB further refines the fused feature $F_{in}$ after multi-scale feature aggregation. This module is designed to mitigate feature truncation and gradient attenuation that may occur in deep networks for sparse point clouds due to the frequent use of explicit activation functions such as Rectified Linear Unit (ReLU) or Gaussian Error Linear Unit (GELU). To improve the stability of feature representations, the AFFRB adopts an activation-free dual-stream residual structure without explicit nonlinear activations. Specifically, the input feature first undergoes layer normalization and multi-scale convolutional processing. It is then fed into a lightweight gating mechanism for activation-free feature interaction. Within this mechanism, the input feature is split into two parts, $X_1$ and $X_2$, along the channel dimension, and a smooth nonlinear transformation is achieved through element-wise multiplication:

$$Y = X_1 \otimes X_2 \tag{4}$$

This linear gating mechanism avoids explicit activation functions while helping preserve feature magnitude information. Subsequently, a simplified Spatial–Channel Attention (SCA) module is introduced to recalibrate the features. The output of the SCA module is combined with the input feature through a residual connection, yielding an intermediate feature $F_{mid}$. The intermediate feature is then fed into a feed-forward network (FFN) path to further enhance feature interactions along the channel dimension. Finally, a residual connection is applied to produce the output feature of the AFFRB:

$$F_{out} = F_{mid} \bigoplus F_{ffn} \tag{5}$$

Through the cascaded design of the AMSAB and AFFRB, the MSFE module is able to recover fine-grained geometric information lost during downsampling to a certain extent, while maintaining stable gradient propagation in deep networks. As a result, the model gains improved capability in capturing complex geometric structures and multi-scale semantic relationships.

### *3.2 Convolution-Attention Mixing Module (CA-Mix)*

Local neighborhood aggregation alone constrains the effective receptive field and limits the capture of non-local semantic dependencies. In complex indoor environments, structurally or semantically correlated regions may not be spatially adjacent, which necessitates explicit non-local interaction. To this end, a Convolution-Attention Mixing (CA-Mix) module is introduced, in which the attention branch explicitly captures non-local contextual dependencies, while the convolution branch preserves local geometric structures. As the non-local component of the proposed non-local multi-scale aggregation framework, CA-Mix complements multi-scale feature enhancement and improves semantic improves segmentation performance in complex scenes.

*3.2.1 Dual-Stream Aggregation*

As illustrated in Fig. 4, the input feature representation is given as $X \in R^{B\times C\times N}$, where B, C, and N denote the batch size, channel dimension, and number of points. The input is first projected through a 3D Depthwise Separable Convolution (DSC) projection to encode preliminary spatial and channel information. The projected features are then processed by a dual-stream aggregation design from two complementary perspectives: global context and local geometry.

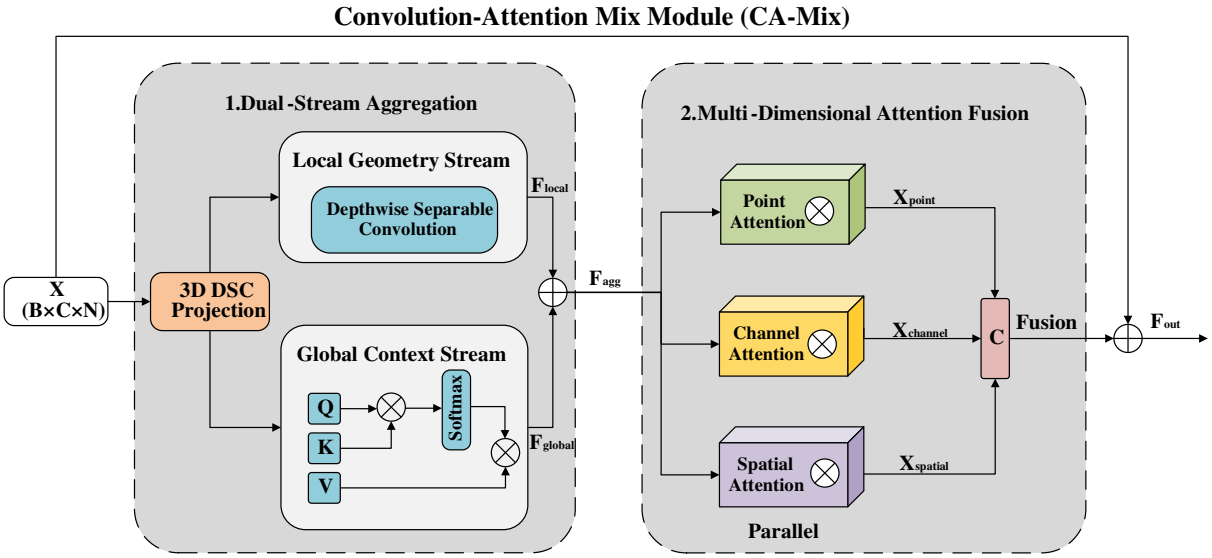


**Figure 4:** Architecture of the proposed Convolution–Attention Mixing (CA-Mix) module.

In the global context branch, the design is inspired by self-attention mechanisms and explicitly introduces non-local interaction across the entire point set. Unlike local neighborhood aggregation, this branch allows each point to establish dependencies with all other points through feature correlation learning, thereby capturing spatially distant but semantically related structures. To avoid the computational overhead of purely attention-based approaches [18,20,21] while preserving local spatial structures in point cloud data, three linear transformations based on the projected features are applied to generate the query $Q$, key $K$, and value $V$ features. Non-local contextual information is then aggregated through scaled dot-product attention, yielding the context-enhanced feature representation $F_{global}$:

$$F_{global} = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

Meanwhile, the local geometry stream applies depthwise separable convolution to the projected features to perform local neighborhood aggregation and structural encoding, producing the local feature representation $F_{local}$. Operating in parallel with the global context stream, it captures fine grained geometric information and complements the non-local interactions introduced by attention. The outputs of the two streams are fused through element wise addition to obtain the aggregated representation $F_{agg}$, which integrates local geometric structures with non-local contextual information.

### *3.2.2 Multi-Dimensional Attention Fusion Block (MDAFB)*

After obtaining the fused feature, a multi-dimensional attention fusion mechanism is further introduced to recalibrate the features from three perspectives: point-level, channel-level, and spatial-level attention. This design aims to suppress redundant responses and highlight discriminative features. Specifically, attention in each dimension modulates the features by generating a corresponding weighting mask $M_k$, and the overall process can be expressed in a unified form as follows:

$$M_k = \sigma\left(\mathcal{W}_k\left(\mathrm{F}_d\right)\right), k \in [\text{point},\ \text{channel},\ \text{spatial}] \tag{7}$$

here, $\mathcal{W}_k(\cdot)$ denotes the weight generation mapping for the corresponding branch.

Point-level attention aims to characterize the relative importance of different points in semantic representation, with a focus on variations in semantic response strength. This branch generates a point-level weighting mask $M_{point} \in R^{B\times 1\times N}$ based on the semantic activation intensity of each point in the fused feature. The mask is then used to modulate the linearly projected features, producing the point-weighted feature $X_{point} = F_{agg} \otimes M_{point}$. By emphasizing points with stronger semantic discrimination, this attention mechanism enhances the representation of key regions in the scene.

Channel-level attention focuses on the contribution differences among feature channels and is used to capture inter-channel dependencies. In this branch, a channel-wise weighting mask $M_{channel} \in R^{B\times C\times 1}$ is generated to recalibrate the fused feature $F_{agg}$ through channel-wise modulation. The reweighted output is given by $X_{channel} = F_{agg} \otimes M_{channel}$.

Spatial-level attention focuses on characterizing the spatial distribution of structural patterns in point cloud data and enhancing responses to geometric boundaries and structural regions. Unlike point-level attention, this branch does not emphasize the semantic saliency of individual points. Instead, it compresses feature information along the channel dimension and performs spatial feature reweighting to generate a structure-aware spatial weighting mask $M_{spatial} \in R^{B\times 1\times N}$. The spatially reweighted feature is then obtained as $X_{spatial} = F_{agg} \otimes M_{spatial}$, which strengthens the representation of object boundaries and local structural regions.

After the three attention-weighted features $X_{point}$, $X_{channel}$, and $X_{spatial}$ are obtained, they are fused by channel-wise concatenation followed by a linear projection. A residual connection is then applied to produce the final output feature:

$$F_{out} = X + \mathcal{F}_{fusion}\left(Concat[X_{point}, X_{channel}, X_{spatial}]\right) \tag{8}$$

With this design, the CA-Mix module builds upon dual-stream feature aggregation and jointly reweights features from three complementary perspectives: point-level semantic responses, channel-wise dependencies, and spatial structural distributions. This coordinated reweighting enhances the network's ability to capture multi-scale semantic and geometric information in complex scenes.

## 4 Experiments

### *4.1 Dataset and Evaluation Metrics*

PointNMSA is evaluated on the S3DIS dataset [9], which is a widely used benchmark for indoor point cloud semantic segmentation. The dataset contains 272 room scans collected from six indoor areas (Area 1–Area 6) and is annotated with 13 semantic categories. Following the standard evaluation protocol, Area 5 is used as the test set, while the remaining areas are used for training. All experimental results are reported on Area 5.

In addition, experiments are conducted on the ScanNet dataset [10], another large scale benchmark for indoor scene understanding. ScanNet consists of richly annotated RGB-D reconstructions of real indoor environments and contains 1513 scanned scenes with 20 semantic categories commonly used for evaluation. Following the official data split, the standard training and validation sets are adopted for performance comparison.

During data preprocessing, point cloud coordinates are first normalized, and voxel-based downsampling is applied to reduce computational cost, with the voxel size set to 0.04. This setting provides a reasonable balance between preserving geometric details and controlling computational complexity, and is consistent with the preprocessing strategy used in the PointNeXt baseline. Instead of block-based cropping, entire

rooms are adopted as network inputs to retain global spatial structure. For fair comparison, all baseline methods follow the same preprocessing pipeline and input configuration.

For performance evaluation, widely accepted metrics in 3D point cloud semantic segmentation are adopted to provide a comprehensive and fair assessment. These metrics include mean Intersection over Union (mIoU), Overall Accuracy (OA), and mean Class Accuracy (mAcc). Specifically, mIoU is computed by averaging the intersection-over-union scores over all semantic categories and reflects overall segmentation quality. OA measures the proportion of correctly classified points, while mAcc computes the average accuracy across all classes, which helps mitigate the impact of class imbalance.

### *4.2 Experimental Setup*

All experiments are conducted on a workstation equipped with an NVIDIA RTX 3090 GPU (24 GB memory), running Ubuntu 20.04. Model training and inference are implemented using PyTorch and accelerated with Compute Unified Device Architecture (CUDA).

During training, the cross-entropy loss is used as the optimization objective, and a label smoothing strategy is applied to improve generalization, with the smoothing factor set to 0.2. The model is trained for 100 epochs with a batch size of 16. The AdamW optimizer is adopted with a weight decay of 0.05, and gradient norm clipping is applied with a maximum value of 28 to stabilize training. The initial learning rate is set to 0.006 and is dynamically adjusted using a cosine annealing schedule to ensure stable convergence throughout training.

For the main ablation results in overall module ablation results on S3DIS Area 5, each configuration was evaluated over three independent runs with different random seeds. The reported results are averaged across these runs, and the mean with standard deviation of mIoU is additionally provided to reflect performance stability, as mIoU is the primary evaluation metric for semantic segmentation. For the remaining ablation and comparison experiments, all results are obtained from a single run with a fixed random seed to ensure fair and consistent comparisons across different methods and configurations.

### *4.3 Ablation Study*

To evaluate the effectiveness of the proposed Multi-Scale Feature Enhancement (MSFE) and Convolution-Attention Mixing (CA-Mix) modules, systematic ablation studies are conducted on the PointNeXt baseline [8]. All experiments are carried out under the same data split, training strategy, and evaluation settings to ensure fair comparison.

As shown in Table 1, progressively introducing MSFE and CA-Mix into the PointNeXt baseline leads to consistent improvements in segmentation performance. When MSFE is added alone, the mIoU increases from 63.45% to 64.50%, while both OA and mAcc also improve slightly. This improvement indicates that multi-scale feature enhancement helps recover geometric details lost during hierarchical downsampling. When CA-Mix is introduced independently, the mIoU increases to 64.82%, suggesting that incorporating non-local contextual interaction improves semantic discrimination in complex indoor scenes. Compared with MSFE, CA-Mix yields a larger performance gain, indicating that it plays a more dominant role in driving the overall improvement, while MSFE mainly provides complementary enhancement.

When both modules are integrated, the proposed PointNMSA achieves the best performance with 65.05% mIoU, 88.78% OA, and 71.30% mAcc. However, the combined improvement is smaller than the sum of their individual gains, indicating that the two modules are not strictly additive. This can be attributed to their partially overlapping effects in structurally complex regions, which limit the accumulated gain. Such overlap is more likely to occur in categories characterized by ambiguous boundaries or fine-grained

structures (e.g., window and board), where both global context and local geometric details are simultaneously required, leading to partially redundant improvements. Similarly, for thin or sparsely represented structures (e.g., column), both modules may respond to the same structural cues, further reducing the independence of their contributions. Therefore, the advantage of combining the two modules lies not in a simple superposition of gains, but in achieving a better balance between geometric detail preservation and contextual semantic discrimination.

**Table 1:** Overall module ablation results on S3DIS Area 5.

| Method | MSFE | CA-Mix | OA (%) | mAcc (%) | mIoU (%) | Params (M) | Latency (ms) |
|---|---|---|---|---|---|---|---|
| PointNeXt | ✗ | ✗ | 87.98 | 70.55 | 63.45 ± 0.18 | 3.16 | 42.24 |
| + MSFE | ✓ | ✗ | 88.61 | 70.82 | 64.50 ± 0.16 | 3.35 | 43.41 |
| + CA-Mix | ✗ | ✓ | 88.65 | 71.05 | 64.82 ± 0.15 | 8.48 | 44.62 |
| **+MSFE + CA-Mix (Ours)** | ✓ | ✓ | **88.78** | **71.30** | **65.05 ± 0.14** | **8.67** | **45.18** |

The relatively small standard deviations reported in Table 1 (ranging from ±0.14 to ±0.18) indicate that the performance variations across independent runs remain limited, suggesting that the observed improvements are consistently achieved rather than caused by random fluctuations.

Fig. 5 further illustrates the performance-efficiency trade-off under different module configurations. As shown in Fig. 5a, the proposed method achieves consistently improved segmentation performance across different module combinations, indicating the effectiveness of the introduced components. Fig. 5b,c further reveal how this performance improvement is associated with increases in parameter count and inference latency, respectively. According to Table 1, incorporating MSFE increases the parameter count slightly from 3.16 to 3.35M, while the latency rises from 42.24 to 43.41 ms, indicating that the additional overhead in both model size and inference time remains limited. In contrast, introducing CA-Mix results in a more significant increase in parameters from 3.16 to 8.48M, while the latency grows from 42.24 to 44.62 ms. Notably, the increase in inference latency is much smaller than the increase in parameter count. Although the parameter growth is noticeable, it mainly comes from the multi-branch design and additional projection and channel mixing layers introduced in the contextual interaction module, which increase model capacity but do not proportionally increase computational intensity during inference. As a result, the growth in actual inference time remains moderate despite the larger model size. When both modules are combined, the parameter count reaches 8.67M and the latency increases to 45.18 ms, representing a moderate overall increase compared to the baseline. Taken together, these results indicate that the proposed design maintains a controlled increase in inference cost while achieving improved segmentation performance, demonstrating a favorable trade-off between model complexity and practical efficiency.

To further analyze the internal design of the MSFE module, additional ablation experiments are conducted, as reported in Table 2. Starting from the baseline model, introducing the Adaptive Multi-Scale Aggregation Block (AMSAB) alone improves the mIoU from 63.51% to 64.09%, indicating that adaptive multi-scale feature fusion helps capture complementary geometric and semantic information from different resolutions. This first step already brings a gain of 0.58 mIoU with only a 0.07M increase in parameters, suggesting that AMSAB provides a favorable balance between accuracy and computational complexity. Therefore, AMSAB can be regarded as the primary contributor within the MSFE module. When the Spatial-Channel Attention (SCA) mechanism is added (A2), the mIoU further increases to 64.31%, suggesting that feature recalibration enhances the discriminative representation of important features. Compared

with A1, A2 yields an additional gain of 0.22 mIoU with another 0.07M parameters, indicating that SCA mainly enhances the discriminative ability of the fused features rather than introducing new information. Finally, incorporating the FFN refinement unit (A3) leads to the full MSFE module and achieves 64.56% mIoU. Relative to A2, the FFN refinement contributes a further 0.25 mIoU improvement with only 0.05M additional parameters, showing that the refinement stage further improves the feature representation. Compared with AMSAB, the contributions of SCA and FFN are more incremental, mainly providing complementary refinement rather than dominating the performance improvement. AMSAB contributes cross scale feature aggregation, SCA enhances feature selection after fusion, and FFN further refines the aggregated representation. The full MSFE module therefore outperforms each partial variant because these components operate in a complementary manner rather than repeating similar functions.

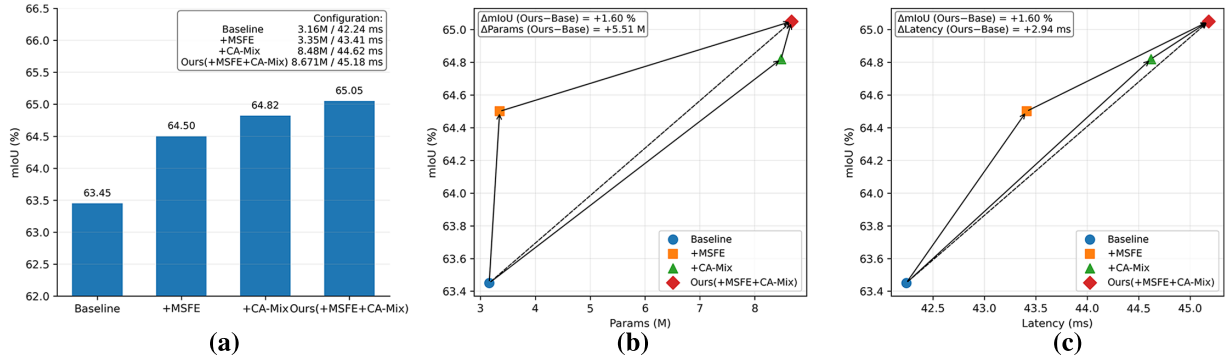


**Figure 5:** Performance-efficiency trade-off analysis under different module configurations: (**a**) comparison of mIoU across different module combinations; (**b**) relationship between parameter count and mIoU; (**c**) relationship between inference latency and mIoU.

**Table 2:** Ablation study on MSFE module.

| ID | AMSAB | SCA | FFN | mIoU | Params (M) |
|---|---|---|---|---|---|
| PointNeXt | ✗ | ✗ | ✗ | 63.51 | 3.16 |
| A1 | ✓ | ✗ | ✗ | 64.09 | 3.23 |
| A2 | ✓ | ✓ | ✗ | 64.31 | 3.30 |
| A3 (Full MSFE) | ✓ | ✓ | ✓ | 64.56 | 3.35 |

Table 3 further investigates the contributions of different components in the CA-Mix module. When only the global attention branch is enabled (B1), the mIoU improves to 64.07%, indicating that non-local contextual interaction helps capture long-range semantic dependencies. The parameter count increases to 5.30M, which can be attributed to the additional projection and interaction operations introduced by the attention mechanism. Using only the local convolution branch (B2) also leads to a slight improvement (63.96% mIoU) with a relatively smaller parameter increase (4.91M), since convolution mainly performs localized feature aggregation without introducing extensive pairwise interactions. When both branches are combined (B3), the mIoU further increases to 64.54%, confirming that global and local feature aggregation provide complementary information, while the parameter count rises to 6.59M due to the joint modeling of local structures and long-range dependencies. This indicates that the integration of global and local branches provides a more balanced representation, but neither branch alone dominates the overall performance. In addition, introducing the Multi-Dimensional Attention Fusion Block (MDAFB) alone (B4) improves

the mIoU to 64.18% with 5.26M parameters, suggesting that feature recalibration enhances discriminative responses with moderate structural overhead. Compared with the global and local branches, MDAFB mainly refines the fused features and plays a complementary role, rather than being the primary driver of performance improvement. When all components are integrated (B5), the full CA-Mix module achieves the best performance of 64.88% mIoU, while the parameter count increases to 8.48M. This increase mainly comes from the combined effect of attention-based interaction, convolutional aggregation, and feature fusion operations. From a performance and model complexity perspective, B1 and B2 provide relatively lightweight improvements, B3 offers a more balanced configuration by integrating global and local representations, and B5 achieves the highest accuracy with increased structural complexity. Overall, the performance gain of CA-Mix is primarily driven by the global attention branch, while the local convolution branch and MDAFB provide complementary and refinement effects. This suggests that the improvement stems not merely from increased model capacity, but from the coordinated interaction among global attention, local convolution, and feature fusion mechanisms.

**Table 3:** Ablation study on CA-Mix module.

| ID | Global | Local | MDAFB | mIoU | Params (M) |
|---|---|---|---|---|---|
| PointNeXt | ✗ | ✗ | ✗ | 63.51 | 3.16 |
| B1 | ✓ | ✗ | ✗ | 64.07 | 5.30 |
| B2 | ✗ | ✓ | ✗ | 63.96 | 4.91 |
| B3 | ✓ | ✓ | ✗ | 64.54 | 6.59 |
| B4 | ✗ | ✗ | ✓ | 64.18 | 5.26 |
| B5(Full CA-Mix) | ✓ | ✓ | ✓ | 64.88 | 8.48 |

### *4.4 Comparative Experiments*

To further evaluate the effectiveness, PointNMSA is compared with several representative point cloud semantic segmentation approaches on the S3DIS Area 5 benchmark. The quantitative comparison results are summarized in Table 4.

**Table 4:** Overall comparison on S3DIS (Area 5).

| Method | OA | mAcc | mIoU |
|---|---|---|---|
| PointNet [2] | 78.9 | 49.0 | 41.1 |
| PointCNN [28] | 85.9 | 63.9 | 57.3 |
| MinkowskiNet20 [5] | – | 69.62 | 62.60 |
| PCM-Tiny [29] | 88.2 | 71.0 | 63.4 |
| PointNeXt (baseline) [8] | 88.03 | 70.69 | 63.51 |
| **Ours (PointNMSA)** | **88.84** | **71.39** | **65.10** |
| KPConv [12] | – | – | 67.1 |
| Point Transformer [18] | 90.8 | 76.5 | 70.4 |

Compared with earlier point-based networks such as PointNet and PointCNN, PointNMSA achieves better segmentation performance. The improvement mainly benefits from the enhanced feature representation introduced by the proposed modules. When compared with several representative methods such as MinkowskiNet and PCM-Tiny, PointNMSA also achieves competitive performance. Although some

methods, such as KPConv and Point Transformer, report higher mIoU values on this benchmark, they adopt more complex convolution operators or transformer-based architectures for feature interaction. In contrast, PointNMSA focuses on enhancing the PointNeXt backbone by incorporating multi-scale feature enhancement and non-local context interaction modules. The consistent improvements over the baseline demonstrate that PointNMSA effectively improves feature representation while maintaining the original architectural framework.

From the per-class Intersection over Union (IoU) results reported in Table 5, the improvements brought by PointNMSA are mainly concentrated on several fine-grained and structurally ambiguous categories. In particular, the IoU gains for window and board reach +8.30 and +8.98, respectively, which contribute most to the overall performance improvement. In addition, categories such as door (+2.23), bookcase (+1.64), and clutter (+2.24) also show consistent positive gains. These categories are often characterized by thin structures, sparse point distributions, or semantic similarity with surrounding objects in indoor scenes. The improvements suggest that the multi-scale feature enhancement and non-local context integration in PointNMSA help improve prediction consistency in regions with ambiguous boundaries and strong semantic confusion. It should also be noted that the IoU of the column category decreases slightly (22.79% → 18.12%). This reduction may be related to the relatively small number of column instances in the dataset and the sensitivity of slender structures to voxel-based downsampling [30]. Despite this localized decline, PointNMSA still achieves improvements in several challenging categories and leads to a higher overall mIoU of 65.10%, compared with 63.51% for the baseline.

**Table 5:** Per-class IoU on S3DIS (Area 5).

| Method | Ceil | Floor | Wall | Beam | Col | Win | Door |
|---|---|---|---|---|---|---|---|
| PointNeXt | 93.70 | 98.40 | 81.36 | 0.0 | 22.79 | 47.68 | 66.31 |
| Ours | 94.72 (+1.02) | 98.36 (−0.04) | 81.97 (+0.61) | 0.0 | 18.12 (−4.67) | 55.98 (+8.30) | 68.54 (+2.23) |
| **Method** | **Table** | **Chair** | **Sofa** | **Book** | **Board** | **Clut** | **mIoU (%)** |
| PointNeXt | 81.38 | 88.80 | 64.86 | 69.54 | 58.00 | 52.78 | 63.51 |
| Ours | 81.26 (−0.12) | 88.49 (−0.31) | 65.73 (+0.87) | 71.18 (+1.64) | 66.98 (+8.98) | 55.02 (+2.24) | 65.10 |

### *4.5 Cross-Dataset Generalization*

To further evaluate the generalization ability of PointNMSA, additional experiments are conducted on the ScanNet dataset. The same model configuration is directly applied to ScanNet without modifying the network architecture. The results are reported in Table 6.

**Table 6:** Cross-dataset generalization results on ScanNet.

| Method | OA (%) | mAcc (%) | mIoU (%) |
|---|---|---|---|
| PointNeXt | 86.35 | 74.54 | 65.25 |
| + MSFE | 87.77 | 75.59 | 67.27 |
| + CA-Mix | 87.66 | 75.43 | 66.82 |
| PointNMSA (MSFE + CA-Mix) | 87.87 | 75.72 | 67.34 |

As shown in Table 6, PointNMSA consistently improves segmentation performance over the PointNeXt baseline on the ScanNet dataset. The mIoU increases from 65.25% to 67.34%, while OA and mAcc also show noticeable improvements. Similar to the results on S3DIS, introducing either MSFE or CA-Mix individually improves performance, and their combination achieves the best results. These results indicate that PointNMSA maintains stable performance across different indoor datasets and demonstrates good cross-dataset generalization ability.

### *4.6 Qualitative Visualization Analysis*

To further examine the segmentation behavior of PointNMSA in complex indoor scenes, qualitative visual comparisons are conducted on two representative rooms from the S3DIS Area 5 test set, as shown in Figs. 6 and 7. The figure presents the ground-truth labels, the predictions of the PointNeXt baseline, and PointNMSA predictions for comparison.

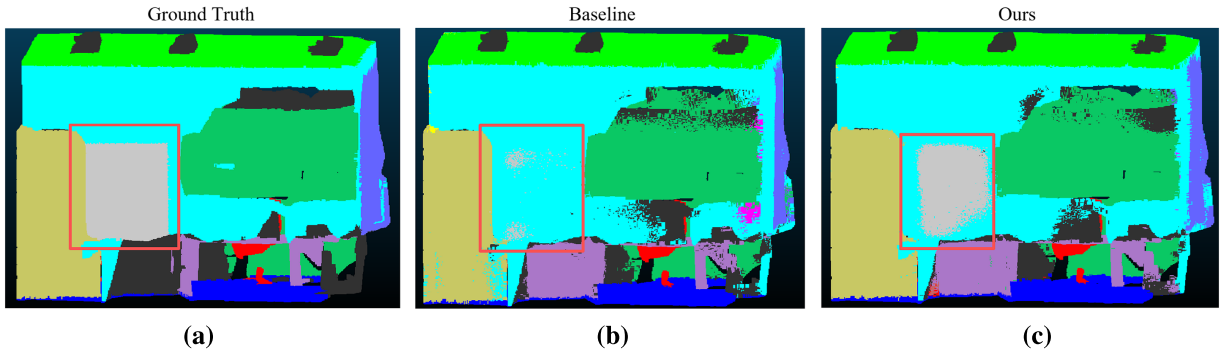


**Figure 6:** Qualitative visualization of segmentation results on a representative S3DIS Area 5 room: (**a**) ground-truth labels; (**b**) predictions of the PointNeXt baseline; (**c**) predictions of PointNMSA.

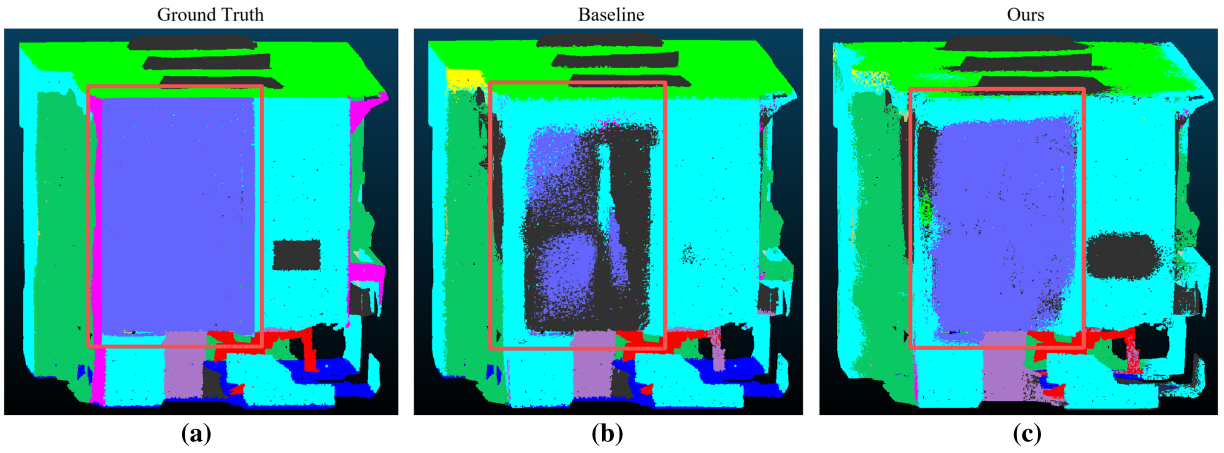


**Figure 7:** Qualitative visualization of segmentation results on another representative S3DIS Area 5 room: (**a**) ground-truth labels; (**b**) predictions of the PointNeXt baseline; (**c**) predictions of PointNMSA.

As shown in Fig. 6a–c, this scene focuses on the segmentation performance for the Board category. Since boards are typically adjacent to walls and exhibit relatively flat geometries with weak boundary cues, the PointNeXt baseline in Fig. 6b tends to produce blurred boundaries and confusion with the Wall category, resulting in partial misclassification of board regions. In contrast, PointNMSA in Fig. 6c more

accurately recovers the overall board structure within the highlighted regions, yielding clearer boundaries and predictions that are more consistent with the ground truth shown in Fig. 6a. These results demonstrate the advantage of the proposed approach in capturing fine-grained planar structures.

As illustrated in Fig. 7a–c, this scene highlights the segmentation results for the Window category. Windows usually exhibit thin structures, sparse point distributions, and geometric similarity to surrounding background regions. Under these conditions, the PointNeXt baseline in Fig. 7b often misclassifies window points as Clutter, leading to noticeable semantic noise and fragmented predictions that degrade structural continuity. By comparison, PointNMSA in Fig. 7c significantly reduces misclassified Clutter points within the window regions, enabling a more complete recovery of window structures in the highlighted areas. This improvement can be attributed to the incorporation of multi-scale feature enhancement and non-local contextual integration, which helps enforce prediction consistency when local geometric cues are limited.

## 5 Conclusion

This paper presents PointNMSA, an improved point cloud semantic segmentation network built upon the PointNeXt. The method aims to enhance the representation of complex indoor scenes by strengthening both multi-scale structural features and non-local contextual interactions. By introducing a Multi-Scale Feature Enhancement (MSFE) module and a Convolution-Attention Mixing (CA-Mix) module, PointNMSA strengthens the joint representation of fine-grained geometric information and non-local semantic context while preserving the efficiency of the original backbone.

The MSFE module alleviates the loss of detailed information during downsampling through cross-level feature fusion in the decoding stage, whereas the CA-Mix module introduces non-local contextual interaction while preserving local geometric structures through a dual-stream aggregation design and multi dimensional feature reweighting. The two modules are complementary and together provide stable performance improvements without introducing excessive inference latency.

Experimental results on the S3DIS Area 5 test set demonstrate that PointNMSA outperforms the PointNeXt baseline in overall segmentation performance as well as on confusing categories such as Board and Window. Additional cross dataset experiments on ScanNet further show that the proposed architecture maintains stable improvements under different indoor scene distributions, indicating good generalization capability. Qualitative visualization further confirms the advantages of the proposed approach in terms of structural integrity and semantic consistency, in agreement with the quantitative results.

In future work, we will further improve the computational efficiency of PointNMSA and enhance its ability to handle slender structures and class imbalance. Robustness under more challenging conditions, such as sparse point density, noise, and occlusion, will also be further investigated.

**Acknowledgement:** Not applicable.

**Funding Statement:** The authors received no specific funding for this study.

**Author Contributions:** The authors confirm contribution to the paper as follows: study conception and design, Aihua Wu and Chenlu Huang; analysis and interpretation of results: Chenlu Huang; draft manuscript preparation: Chenlu Huang; manuscript revision: Aihua Wu and Chenlu Huang. All authors reviewed and approved the final version of the manuscript.

**Availability of Data and Materials:** The datasets used in this study are publicly available.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References


1. Schnabel R, Wahl R, Klein R. Efficient RANSAC for point-cloud shape detection. Comput Graph Forum. 2007;26(2):214–26. doi:10.1111/j.1467-8659.2007.01016.x.
2. Charles RQ, Su H, Kaichun M, Guibas LJ. PointNet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 77–85. doi:10.1109/cvpr.2017.16.
3. Qi CR, Yi L, Su H, Guibas LJ. Pointnet++: deep hierarchical feature learning on point sets in a metric space. Adv Neural Inf Process Syst. 2017;30:5099–108.
4. Zhou Y, Tuzel O. VoxelNet: end-to-end learning for point cloud based 3D object detection. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 4490–9. doi:10.1109/cvpr.2018.00472.
5. Choy C, Gwak J, Savarese S. 4D spatio-temporal ConvNets: minkowski convolutional neural networks. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 3070–9. doi:10.1109/cvpr.2019.00319.
6. Phan AV, Le Nguyen M, Nguyen YLH, Bui LT. DGCNN: a convolutional neural network over large-scale labeled graphs. Neural Netw. 2018;108(4):533–43. doi:10.1016/j.neunet.2018.09.001.
7. Sarker S, Sarker P, Stone G, Gorman R, Tavakkoli A, Bebis G, et al. A comprehensive overview of deep learning techniques for 3D point cloud classification and semantic segmentation. Mach Vis Appl. 2024;35(4):67. doi:10.1007/s00138-024-01543-1.
8. Qian G, Li Y, Peng H, Mai J, Hammoud HAAK, Elhoseiny M, et al. PointNeXt: revisiting PointNet++ with improved training and scaling strategies. In: Proceedings of the Advances in Neural Information Processing Systems 35; 2022 Nov 28–Dec 9; New Orleans, LA, USA. p. 23192–204. doi:10.52202/068431-1685.
9. Armeni I, Sax S, Zamir A, Savarese S. Joint 2D–3D–semantic data for indoor scene understanding. arXiv:1702.01105. 2017.
10. Dai A, Chang AX, Savva M, Halber M, Funkhouser T, Niessner M. ScanNet: richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 2432–43. doi:10.1109/cvpr.2017.261.
11. Chang A, Dai A, Funkhouser T, Halber M, Niebner M, Savva M, et al. Matterport3D: learning from RGB-D data in indoor environments. In: Proceedings of the 2017 International Conference on 3D Vision (3DV); 2017 Oct 10–12; Qingdao, China. p. 667–76. doi:10.1109/3dv.2017.00081.
12. Thomas H, Qi CR, Deschaud JE, Marcotegui B, Goulette F, Guibas L. KPConv: flexible and deformable convolution for point clouds. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 6410–9. doi:10.1109/iccv.2019.00651.
13. Wu W, Qi Z, Li F. PointConv: deep convolutional networks on 3D point clouds. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 9613–22. doi:10.1109/cvpr.2019.00985.
14. Li H, Guan H, Ma L, Lei X, Yu Y, Wang H, et al. MVPNet: a multi-scale voxel-point adaptive fusion network for point cloud semantic segmentation in urban scenes. Int J Appl Earth Obs Geoinf. 2023;122(12):103391. doi:10.1016/j.jag.2023.103391.
15. Wang X, Cui K, Wang L, Liu Z, Yu B, He Y, et al. VPFNET: a scale-adaptive voxel point fusion network for Semantic segmentation of point clouds. In: Proceedings of the Pattern Recognition and Computer Vision—PRCV 2024; 2024 Oct 18–20; Urumqi, China. p. 75–88. doi:10.1007/978-981-97-8792-0_6.
16. Hu Q, Yang B, Xie L, Rosa S, Guo Y, Wang Z, et al. RandLA-net: efficient semantic segmentation of large-scale point clouds. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 11105–14. doi:10.1109/cvpr42600.2020.01112.
17. Zeng Z, Hu Q, Xie Z, Li B, Zhou J, Xu Y. Small but mighty: enhancing 3D point clouds semantic segmentation with U-Next framework. Int J Appl Earth Obs Geoinf. 2025;136(11):104309. doi:10.1016/j.jag.2024.104309.

18. Zhao H, Jiang L, Jia J, Torr P, Koltun V. Point transformer. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 16239–48. doi:10.1109/iccv48922.2021.01595.
19. Guo MH, Cai JX, Liu ZN, Mu TJ, Martin RR, Hu SM. PCT: point cloud transformer. Comp Visual Med. 2021;7(2):187–99. doi:10.1007/s41095-021-0229-5.
20. Lai X, Liu J, Jiang L, Wang L, Zhao H, Liu S, et al. Stratified transformer for 3D point cloud segmentation. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 8490–9. doi:10.1109/cvpr52688.2022.00831.
21. Yang YQ, Guo YX, Xiong JY, Liu Y, Pan H, Wang PS, et al. Swin3D: a pretrained transformer backbone for 3D indoor scene understanding. Comp Visual Med. 2025;11(1):83–101. doi:10.26599/cvm.2025.9450383.
22. Yang YQ, Guo YX, Liu Y. Swin3D++: effective multi-source pretraining for 3D indoor scene understanding. Comp Visual Med. 2025;11(3):465–81. doi:10.26599/cvm.2025.9450437.
23. Liu S, Chi J, Wu C, Xu F, Yu X. SGT-net: a transformer-based stratified graph convolutional network for 3D point cloud semantic segmentation. Comput Mater Contin. 2024;79(3):4471–89. doi:10.32604/cmc.2024.049450.
24. Zhang S, Wang B, Chen Y, Zhang S, Zhang W. Point and voxel cross perception with lightweight cosformer for large-scale point cloud semantic segmentation. Int J Appl Earth Obs Geoinf. 2024;131:103951. doi:10.1016/j.jag.2024.103951.
25. Wu X, Jiang L, Wang PS, Liu Z, Liu X, Qiao Y. Point transformer V3: simpler, faster, stronger. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 4840–51. doi:10.1109/cvpr52733.2024.00463.
26. Robert D, Raguet H, Landrieu L editors. Efficient 3D semantic segmentation with superpoint transformer. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 17149–58. doi:10.1109/iccv51070.2023.01577.
27. Shu J, Wang S, Yu S, Zhang J. CFSA-Net: efficient large-scale point cloud semantic segmentation based on cross-fusion self-attention. Comput Mater Contin. 2023;77(3):2677–97. doi:10.32604/cmc.2023.045818.
28. Li Y, Bu R, Sun M, Wu W, Di X, Chen B. PointCNN: convolution on x-transformed points. Adv Neural Inf Process Syst. 2018;31:828–38.
29. Zhang T, Yuan H, Qi L, Zhang J, Zhou Q, Ji S, et al. Point cloud mamba: point cloud learning via state space model. Proc AAAI Conf Artif Intell. 2025;39(10):10121–30. doi:10.1609/aaai.v39i10.33098.
30. Li M, Lin S, Wang Z, Shen Y, Zhang B, Ma L. Class-imbalanced semi-supervised learning for large-scale point cloud semantic segmentation via decoupling optimization. Pattern Recognit. 2024;156(8):110701. doi:10.1016/j.patcog.2024.110701.