GASegNet: Global Self-Attention Mechanism Meets Structural Feature Fusion for Point Cloud Semantic Segmentation

Xu Lu, Haijun Liu, Guang'an Luo, Zhike Chen, Cheng Zhou, Xinyu Wu, *Senior Member, IEEE*, and Jun Liu

***Abstract*—With the rapid development of autonomous driving technology, semantic segmentation, as one of the key technologies contributing to the environmental perception of autonomous driving systems, still suffers from a lack of connections between local features, as well as high computational consumption and an inability to meet real-time requirements. To address the above problems, this paper proposes a lightweight and efficient point cloud semantic segmentation network based on spherical projection with an encoder-decoder structure. The encoder consists of a global self-attention mechanism that captures global information, as well as multi-scale convolution. This module achieves the unification of local feature extraction and global characteristic information for high-dimensional semantic information. In order to alleviate the high computational cost, a feature fusion module is introduced to enhance the compactness of the range image structure obtained from point cloud projection. The decoder utilizes bilinear interpolation to upsample multi-resolution feature maps and introduces multiple auxiliary segmentation heads to further enhance the network's accuracy. Experiments conducted on the SemanticKITTI and SemanticPOSS datasets reveal that, in comparison to the CENet architecture, the proposed approach attains enhancements in mIoU of 4.3% and 2.6% on the respective datasets, thereby substantiating its efficacy. The code is available at GitHub: https://github.com/haifeng925/GASegNet.**

Link to graphical and video abstracts, and to code:
https://latamt.ieeer9.org/index.php/transactions/article/view/10124

***Index Terms*—Autonomous driving, Spherical projection, Point cloud semantic segmentation, Self-attention mechanism**

I. Introduction

The associate editor coordinating the review of this manuscript and approving it for publication was Martin Pedemonte (*Corresponding author: Jun Liu*).

This work was supported by the Original Exploration Program of the National Natural Science Foundation of China (Grant No. 62550171); National Key Research and Development Program of China (2025YFE0103200); Key-Area Research and Development Program of Guangdong Province (2023B0303020001); Scientific and Technological Planning Project of Guangzhou (2023B03J1378); Research project of Guangdong Polytechnic Normal University (22GPNUZDJS14).

X. Lu is with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China; Guangdong Provincial Key Laboratory of Intellectual Property & Big Data, Guangzhou 510665, China; Pazhou Lab, Guangzhou 510330, China; (e-mail: xulu@gpnu.edu.cn).

Jun Liu is with the School of Automation, Guangdong Polytechnic Normal University, Guangzhou 510665, China; (e-mail: liujun7700@163.com).

X. Wu is with the Shenzhen Institute of Advanced Technology, Shenzhen 518055, China.(e-mail: xy.wu@siat.ac.cn).

H. Liu, G. Luo, Z. Chen, and C. Zhou are with the Guangdong Polytechnic Normal University, Guangzhou 510665, China; (e-mails: haifeng@stu.gpnu.edu.cn, luoguangan@gpnu.edu.cn, chuck@gpnu.edu.cn, and zhoucheng@gpnu.edu.cn).

ARTIFICIAL Intelligence (AI) is widely applied in various fields, including autonomous driving, robot navigation, map building, AR virtual interaction, and industrial inspection. Semantic segmentation helps vehicles understand the semantic information of their surrounding environment, enabling them to make informed decisions.

Semantic segmentation technologies can be categorized into image-based and LiDAR-based approaches. Image semantic segmentation, primarily based on deep learning methods such as convolutional neural networks (CNNs), has seen rapid development. However, its performance degrades significantly under challenging conditions such as strong exposure, insufficient light, or adverse weather, and it lacks accurate spatial information. In contrast, LiDAR technology can obtain accurate 3D spatial information and object geometry by reflecting pulsed laser light. Point cloud semantic segmentation enhances the understanding of self-driving cars by categorizing each 3D point into specific categories. Although early development was limited by the limited availability of datasets, the release of benchmarks such as SemanticKITTI [1] and SemanticPOSS [2] has significantly accelerated progress in this area.

Early point cloud semantic segmentation methods were predominantly point-based, with PointNet [3] being the first pioneering framework. It employs a shared-parameter multilayer perceptron (MLP) to extract individual point features and merges global features using a symmetric function. However, PointNet overlooks local structural information and fails to effectively fuse local and global features, limiting its ability to model fine-grained details and occluded structures.

To overcome these limitations, PointNet++ [4] introduces hierarchical feature learning inspired by CNNs, enabling local neighborhood feature extraction. Nevertheless, it still relies on partial sampling and first-order features, neglecting more detailed and high-dimensional information, and struggles to capture complex geometric relationships and long-range dependencies. Subsequent methods incorporate attention mechanisms to enhance feature modeling. PointTransformer [5] employs self-attention to capture long-range dependencies, while Graph Attention Networks (GAT) [6] and non-local networks [7] further improve contextual modeling. Despite their performance gains, these methods incur high computational costs, which restrict their applicability in real-time scenarios.

RangeNet++ [8] introduces spherical projection to transform 3D point clouds into 2D distance images, enabling the use of image-based segmentation networks. However, due

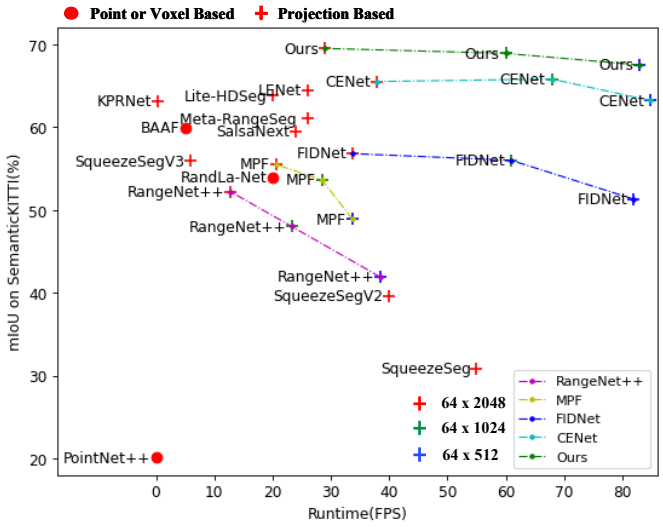

Fig. 1. Accuracy (mIoU) vs. Runtime(FPS) of Segmentation on the SemanticKITTI Validation Set.

to the disordered structure of point cloud data, it achieves only 41.9% mIoU on the SemanticKITTI dataset with 50M model parameters, highlighting the issue of excessive model complexity. SqueezeSegV3 [9] further points out that feature distributions vary significantly across spatial locations after projection, making conventional image segmentation models less effective. To address this issue, SqueezeSegV3 introduces Spatial Adaptive Convolution (SAC), which applies position-dependent filters with an attention-based design. On the SemanticKITTI benchmark, SqueezeSegV3 outperforms recent leading approaches [10] at the time by at least 3.7% mIoU under comparable inference speed, with a reduced model size of 29.9M parameters.

To further reduce model complexity, FIDNet [11] and CENet [12] replace decoder convolutions with bilinear interpolation and avoid using attention mechanisms, significantly reducing the number of model parameters. As a result, segmentation accuracy (mIoU) improves from 58.6% to 65.5%, while the number of model parameters is reduced to 6.783M.

The model proposed in this paper achieves 69.3% mIoU on the SemanticKITTI dataset (shown in Fig. 1), with only a 0.7% increase in model parameters. On the SemanticPOSS dataset, it achieves 52.7% mIoU, representing a 2.4% improvement over the benchmark model CENet. This method effectively addresses the challenges of local structural feature extraction and global contextual information capture while controlling model complexity, thereby improving the accuracy of semantic segmentation models.

The main contributions of this paper are as follows:

- We propose GASegNet, a semantic segmentation network for LiDAR, which incorporates a Feature Fusion Module (FFM) and a Global Self-Attention Mechanism (GSAM) to enhance point cloud semantic segmentation.
- The FFM is introduced to reconstruct geometric structural features of point cloud data and enhance the local feature aggregation capability of the network. This module effectively addresses the challenge of capturing fine-grained details in complex point cloud structures.
- The GSAM is designed to focus on global information, enabling the network to capture higher-level semantic features. By capturing long-distance dependencies between any two locations, GSAM enhances the utilization of global contextual information and addresses the occlusion problem in point cloud data.

II. Related Work

The task of point cloud segmentation for autonomous driving scenarios is divided into three main categories of implementations: point-based methods, voxel-based methods, and image-based methods.

Point-based methods: The methods directly process raw 3D point clouds, preserving spatial structure without information loss. To address the limitations of PointNet [3] in handling multi-scale geometric structures, PointNet++ [4] introduces multi-scale grouping to enhance adaptability. Subsequently, KPConv [13] proposes deformable convolutions with flexible kernel points to better capture local geometric features. BAAF [14] employs bilateral structures and adaptive fusion to jointly exploit geometric and semantic features. DGCNN [15] preserves local geometric structures by constructing local neighborhood graphs using EdgeConv. However, it primarily focuses on learning local features in the feature space and lacks the ability to capture global features, which limits its adaptability to complex point cloud structures. GAPointNet [16] integrates graph attention into stacked MLPs to enhance local geometric feature extraction and robustness. Nevertheless, due to the large scale of point clouds, existing point-based methods still face challenges in high-dimensional feature learning, effective local feature aggregation, and shape information mining. Moreover, high computational cost and limited inter-feature connectivity hinder real-time applicability.

Voxel-based methods: Converting point clouds into voxels for processing can effectively address the unstructured nature of point-cloud information. VoxNet [17] pioneers voxelization to convert unstructured point clouds into regular voxels and applies 3D CNNs for semantic prediction, but suffers from inefficient voxel alignment. Kd-Net [18] improved efficiency by converting point clouds into binary trees and computing only on non-empty voxels. Su et al. proposed SPLATNet [19], which represents point clouds using sparse voxel structures and fuses multi-view features to alleviate the irregularity and sparsity of point clouds. Cylinder3D [20] used an asymmetric residual block to dynamically adjust the feeling field based on point cloud sparsity, reducing the computational effort. AF2S3Net [21] applied attention-based multi-scale feature adaptive fusion without complex hierarchical stacking. However, voxel-based methods face a trade-off: large voxels lose detail, while small voxels increase computational and memory costs, limiting scalability on large-scale point clouds.

Image-based methods: By projecting the point clouds onto multiple views to reduce data dimensionality and simplify 3D structures, enhancing object recognition, classification, and segmentation. SqueezeSeg [22] pioneered this approach for road object segmentation, followed by SqueezeSegV2 [23],

which introduced a context aggregation module, and SqueezeSegV3 [9], which leveraged spatially adaptive convolution to handle feature distribution shifts. RangeNet++ [8] applied accelerated KNN for post-processing, while SalsaNext [24] incorporated uncertainty-aware mechanisms for point feature learning based on SalsaNet [25]. Additionally, KPRNet [26] innovatively replaces the decoder with a KPConv layer, avoiding the over- and under-smoothing problems associated with it. Lite-HDSeg [27] introduces an HD volume operator with a multi-class spatial propagation network to enhance contextual understanding, but struggles with small-object segmentation. FIDNet [11] proposes a new LiDAR semantic segmentation network based on range images. While fully parameter-free, it lacks feature learning capability, leading to suboptimal accuracy due to overreliance on the classification head for model performance and contextual information integration.

In this paper, we propose GASegNet, a lightweight semantic segmentation network based on range images and built upon CENet [12]. To enhance local-global feature interaction while controlling model complexity, we propose a global self-attention mechanism and a novel Feature Fusion Module (FFM). FFM improves structural feature learning and mitigates the impact of uneven point cloud distribution on performance. The attention mechanism captures higher-level semantic information across spatial and channel dimensions, enabling high segmentation accuracy and efficiency with low parameter overhead. The framework is detailed in the following section.

III. Methodology

A. Network Architecture

The network architecture proposed in this paper is shown in Fig. 2. Our method is built upon CENet [12] as the baseline backbone, and extends it by introducing an FFM and a GSAM. Specifically, the 3D point cloud is first converted into a range image by spherical projection. FFM is used to make the range image representation more compact and homogeneous by removing spatial and channel redundancy, which is used to solve the problems of unstructured and inhomogeneous point clouds. The Encoder-Decoder structure is added to the globalization attention mechanism. The Encoder extracts hierarchical semantic features from the projected range image. The Decoder part follows the previous bilinear interpolation method, and a linear branch is induced from the Encoder for obtaining the global consistency of the semantic information. Finally, the KNN algorithm can be used to back-project the segmented image onto the point cloud map. In this structural diagram, only the processing of the image is shown. The step of backprojection onto the point cloud map is placed as a visualization result in Fig. 5. The subsequent subsections detail the Spherical Projection, FFM, GSAM, and loss function.

B. Spherical Projection

The utilization of a spherical projection method to transform point-cloud information into a two-dimensional image representation converts irregular 3D data into a structured 2D grid format, consequently mitigating the structural complexity associated with data processing. This approach enables the application of established two-dimensional image processing techniques and deep learning models for training and analysis. In the range image representation at time t, the (x_t, y_t, z_t) of each LiDAR point p is a Cartesian coordinate, which is converted to an image coordinate (u_t,v_t) using a spherical projection $\mathbb{R}^3 \to \mathbb{R}^2$ as shown in the following equation:

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left[1 - \operatorname{atan2}(y_t, x_t)\pi^{-1}\right] W \\ \left[1 - \left(\arcsin\left(z_t d_t^{-1}\right) + f_u\right)\frac{1}{f_u + f_d}\right] H \end{pmatrix}. \quad (1)$$

The point cloud is projected onto a range image of size $(H, W, 5)$, where H and W denote the fixed height and width of the image, respectively. Each pixel encodes five channels $(x_t, y_t, z_t, d_t, r_t)$, where $d_t = \sqrt{x_t^2 + y_t^2 + z_t^2}$ represents the depth of a point and r_t denotes its reflection intensity. The intensity channel is retained to allow the network to implicitly learn useful cues from material and surface properties. The image coordinates (u_t, v_t) are computed from the 3D point (x_t, y_t, z_t) using the azimuth and elevation angles, as defined in Eq. (1). Here, f_u and f_d denote the upper and lower vertical field-of-view (FoV) angles of the LiDAR sensor, respectively, and the total vertical FoV is given by $f = f_u + f_d$. By reformulating point cloud semantic segmentation as an image-based segmentation problem, this representation enables effective exploitation of both local geometric cues and global contextual information, thereby improving feature learning efficiency.

C. Feature Fusion Module

Due to the unstructured and inhomogeneous problem of point-cloud information, the point cloud will have empty pixels in the middle of the range image projected into it, etc., which introduces a large number of empty or low-information pixels in the projected range image, leading to redundant spatial and channel-wise computations and reduced computational efficiency in semantic segmentation models. Inspired by SCConv [28] to solve the problem, we added a feature fusion module to the CENet [12] network, resulting in a more compact image representation and homogenization of pixel values. As shown in Fig. 3, the FFM is divided into a spatial fusion module and a channel fusion module. The range image data converted from the point cloud is processed by the FFM for fusion on structural features, as well as removing spatial and channel redundancy.

For the input range image $X \in \mathbb{R}^{n \times c \times h \times w}$, where n is the number of batches, c is the number of channels, h and w are the height and width of the image. The feature maps produced by the preceding convolutional layers are normalized using batch normalization (BN) as follows:

$$W = BN(X) = \gamma \frac{X - \mu}{\sqrt{\sigma^2 + \varepsilon}}, \quad (2)$$

where μ and σ are the mean and standard deviation of X, ε is a constant added for stability, and γ is a trainable parameter. The output W of the weights obtained through the standard layer is mapped to the range of (0, 1) through a sigmoid function, and then multiplied by the input range image X. Structurally

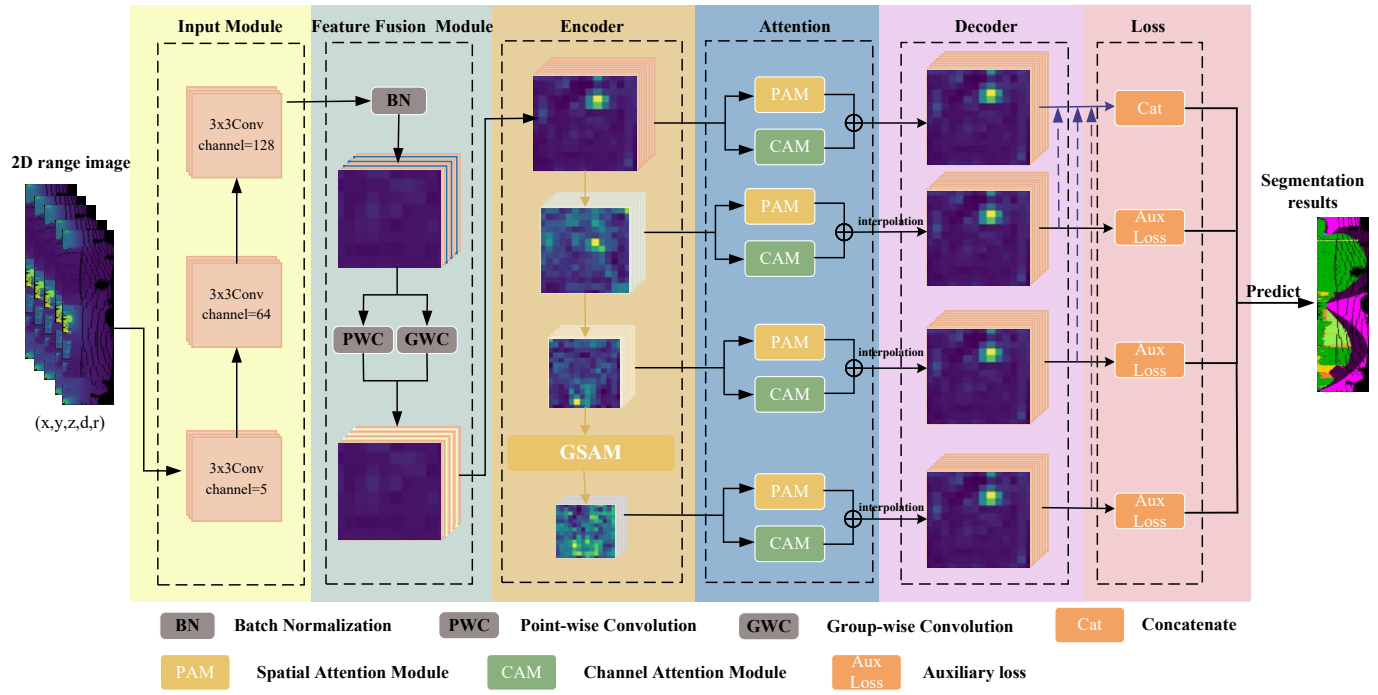

Fig. 2. Network Architecture. We introduce Feature Fusion Module(FFM) and Global Self-Attention Mechanism(GSAM) to enhance the segmentation performance of the network, while optimizing this result with global information and auxiliary loss.

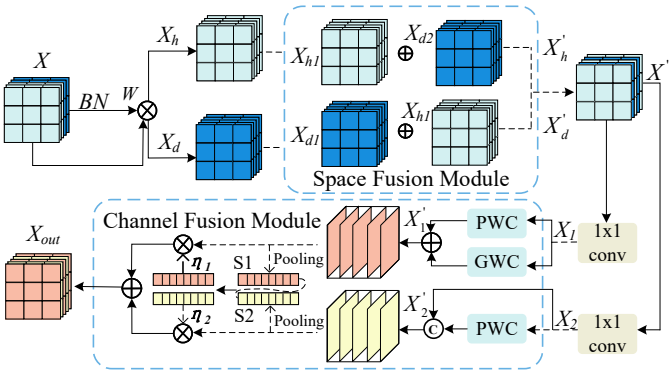

Fig. 3. Feature Fusion Module is divided into two modules.

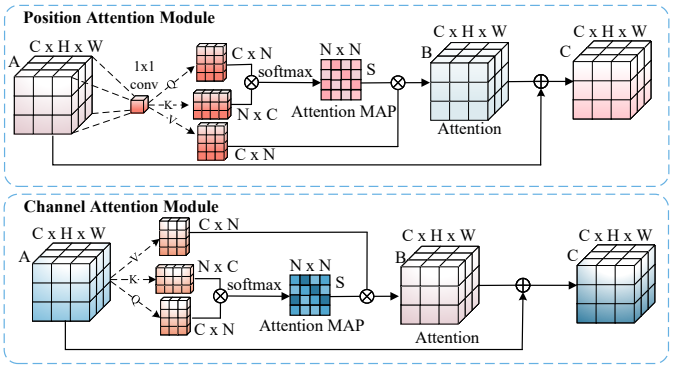

Fig. 4. Global self-attention mechanism with spatial and channel branches. Spatial attention captures long-range dependencies, while channel attention learns channel importance. The fused output produces globally enhanced features.

informative regions in the normalized range image are filtered using a threshold ($X_{\text{threshold}} = 0.5$) to separate informative regions (X_h) from less informative regions (X_d):

$$\begin{cases} X_h, \text{if sigmoid}(WX) \geq X_{\text{threshold}}, \\ X_d, \text{if sigmoid}(WX) < X_{\text{threshold}}. \end{cases} \tag{3}$$

Each subset is further divided along the vertical axis of the range image, which naturally corresponds to different elevation regions in LiDAR scans, denoted as X_{h1}, X_{h2}, X_{d1}, and X_{d2}. The cross-splicing operation is performed on the above two structural features by adding one end of the information-rich feature with one end of the less information-rich feature, and the cross-splicing operation is used to fully combine the weighted two different information features to make the structural features of the data more homogeneous. Then the cross reconstructed features X'_h and X'_d are spliced to get the feature mapping X', the splicing operation is shown below:

$$\begin{cases} X_{h1} \oplus X_{d2} = X'_h, \\ X_{d1} \oplus X_{h2} = X'_d, \\ X'_h \bigcup X'_d = X'. \end{cases} \tag{4}$$

The operation of removing channel redundancy is carried out, and 1x1 convolution is utilized to compress the channels of the feature mapping to improve computational efficiency. We divide the spatial feature mapping X' into X_1 and X_2, and perform GWC [29] and PWC [30] on X_1 to get $X_1^{'}$, the approach was first proposed in Mobilenets [31], and perform PWC on X_2 and adding with itself to get $X_2^{'}$. This process is defined by the following equations, where C^G and C^{P1} denote

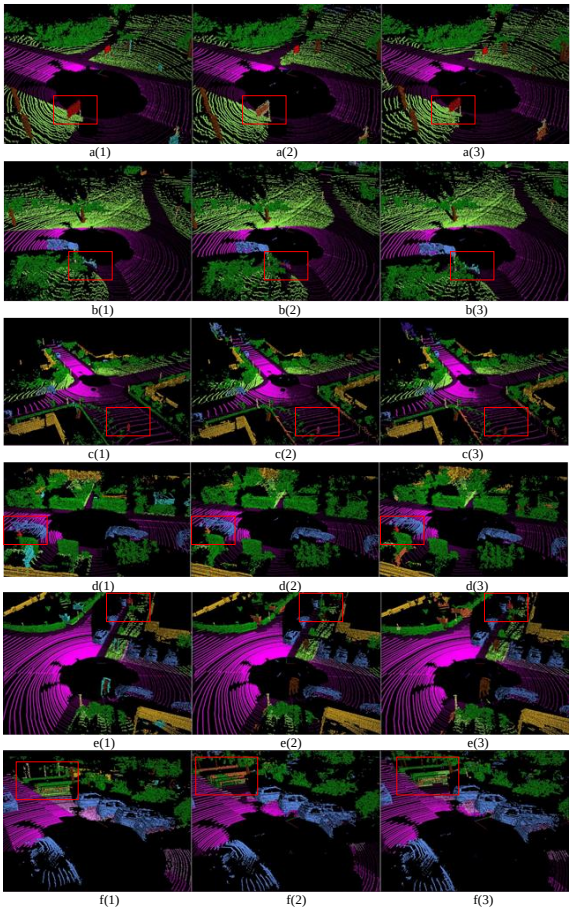

Fig. 5. Qualitative analysis of the SemanticKITTI validation set. Where (1) is the semantic truth value of LiDAR scanned frames, (2) is the semantic graph segmented by the CENet model, and (3) is the semantic graph segmented by our model, GASegNet.

the group and pointwise convolutions for X_1, respectively, and C^{P2} denotes the pointwise convolution for X_2:

$$X_1^{'} = C^G X_1 + C^{P1} X_1, \tag{5}$$

$$X_2^{'} = C^{P2} X_2 \cup X_2. \tag{6}$$

The global average pooling operation is used to collect global spatial information, producing channel descriptor tensors $S_k \in \mathbb{R}^{C\times 1\times 1}$, where $S_k = \frac{1}{h\times w}\sum^j \sum^i X_k^{'}(i,j), k = 1,2$. The scalar descriptors s_k are obtained by reducing S_k along the channel dimension and are used to compute the attention weights η_1 and η_2 as follows:

$$\eta_1 = \frac{e^{s_1}}{e^{s_1}+e^{s_2}}, \eta_2 = \frac{e^{s_2}}{e^{s_1}+e^{s_2}}, \eta_1+\eta_2 = 1. \tag{7}$$

Finally, under the effect of feature vectors η_1, η_2, the upper feature $X_1^{'}$ and lower feature $X_2^{'}$ are merged in the channel direction to obtain the channel-compressed feature X_{out} , which is calculated as follows:

$$X_{out} = \eta_1 X_1^{'} + \eta_2 X_2^{'}. \tag{8}$$

D. Global Self-Attention Mechanism

To address the inaccuracy of semantic segmentation networks caused by occlusion in point cloud information, we propose a lightweight global self-attention mechanism module that captures global context. Additionally, a dual-branch structure is designed to fuse local and global features by mining spatial and channel dimensions, enhancing the network's semantic representation and segmentation capability.

The global self-attention mechanism module consists of two parallel components: a spatial attention module and a channel attention module. A block diagram of this module is shown in Fig. 4. The input feature map undergoes three 3×3 Conv operations. By stacking multiple 3×3 convolutions, the model achieves a smaller number of parameters compared to 7×7 convolutions while gradually expanding the receptive field to capture a larger range of contextual information. This enables the network to extract deeper features and capture contextual information. The spatial and channel attention modules operate in parallel: the spatial attention module captures inter-pixel long-distance dependencies, while the channel attention module extracts global features through global pooling, enhancing the weights of important channels. By fusing the outputs of these two modules, the network generates a feature map containing global context information with attention weights. Unlike convolution, which aggregates local neighborhoods with fixed weights, the proposed attention mechanism dynamically assigns weights based on global feature similarity.

The input feature map $A \in \mathbb{R}^{C\times H\times W}$ is passed through three independent convolutional layers to generate query (Q), key (K), and value (V) $Q, K, V \in \mathbb{R}^{C\times H\times W}$ feature maps. To compute the similarity matrix, Q and K are reshap $\mathbb{R}^{C\times N}$, where C represents the number of channels and N represents the number of pixels. The similarity matrix $S \in \mathbb{R}^{C\times C}$ is then calculated as the dot product of Q and K, capturing global spatial dependencies by measuring pixel-wise similarity. This formulation follows the standard self-attention paradigm, where similarity-based weighting enables long-range dependency modeling:

$$S_{ij} = \frac{\exp(Q_i \cdot K_j)}{\sum_{j=1}^{C}\exp(Q_i \cdot K_j)}. \tag{9}$$

The similarity matrix S is multiplied by the reshaped features of V using matrix multiplication to obtain enhanced features, denoted as B. These enhanced features are then added to the original feature map A to produce C, which aggregates local and global contextual information. The channel attention module follows a similar process, applying attention weights to the feature channels to capture global information and enhance the representation of important features:

$$B_j = \sum_{i=1}^{C} S_{ij} \cdot A_i, \tag{10}$$

$$C_j = A_j + B_j. \tag{11}$$

E. Loss Function

To address three prevalent challenges [12], [32]–[34] in point cloud semantic segmentation: class imbalance, optimized

intersection over union (IoU), and ambiguous segmentation boundaries, we propose the integration of three loss functions during network training: $\mathcal{L}_{wce}$, $\mathcal{L}_{ls}$, and $\mathcal{L}_{bd}$. Here, $\mathcal{L}_{wce}$ denotes the weighted cross-entropy loss, $\mathcal{L}_{ls}$ denotes the Lovász-Softmax loss, a differentiable surrogate of the IoU that applies the Lovász extension to sorted prediction errors, enabling direct optimization of mIoU. Finally, $\mathcal{L}_{bd}$ denotes the boundary-aware loss. Furthermore, we introduce an auxiliary loss mechanism $L(y_i, \hat{y}_i)$, which utilizes the dependency relationships between the three outputs derived from the bilinear interpolation module and the ground truth to enhance network convergence. The total loss function is defined as follows:

$$\mathcal{L} = w_1\mathcal{L}_{wce} + w_2\mathcal{L}_{1s} + w_3\mathcal{L}_{bd} + \sum_{i=1}^{3} L(y_i, \hat{y}_i), \qquad (12)$$

where w_1, w_2 and w_3 represent the weights assigned to each loss function. Based on empirical observations, we set the weights w_1, w_2 and w_3 to 1.25. $\hat{y}_i$ denotes the semantic output at bilinear interpolation module i, and y_i corresponds to the truth semantic label. The auxiliary loss is defined as $\sum_{i=1}^{3} L(y_i, \hat{y}_i)$.

IV. Experiments

A. Dataset and Implementation Details

SemanticKITTI [1] dataset is a widely used benchmark for 3D semantic segmentation in autonomous driving. It is derived from the LiDAR scans of the KITTI Vision Benchmark Suite, with each point cloud carefully annotated into 25 semantic classes, including roads, vehicles, pedestrians, and buildings, with 19 classes commonly used for evaluation.

The SemanticPOSS [2] dataset, a collaborative effort between the Beijing Institute of Technology and Tsinghua University, represents a high-quality LiDAR dataset specifically designed for semantic segmentation tasks in campus environments. Characterized by its extensive scene diversity and meticulously designed annotation system, this dataset holds substantial research significance, particularly in advancing studies related to autonomous driving.

For data acquisition, the research team utilized a Velodyne 16-line LiDAR system to ensure precision and reliability of the collected data. Structurally, the dataset is organized into six continuous sequences, comprising approximately 3000 frames of point cloud data. It encompasses 14 distinct semantic categories, including but not limited to ground, buildings, trees, and pedestrians. The dataset contains a total of 2988 complex LiDAR scan samples, capturing a wide array of scenarios such as roadways in academic zones, residential complexes, and vegetated areas. These scenes incorporate numerous dynamic elements, such as moving pedestrians and vehicles, as well as real-world challenges like occlusions and dynamic targets.

Implementation details: In this paper, the proposed method is implemented using OpenPCSeg Codebase [35]. We implemented the proposed method using PyTorch on the NVIDIA RTX2080 platform, and during the training process, we used data enhancement to perform preprocessing operations on the data. We employed the AdamW [36] optimizer with default settings in PyTorch. The weight decay was set to $1e^{-4}$. We trained the network using an initial learning rate of $1e^{-2}$ and dynamically adjusted the learning rate over 100 epochs.

B. Results and Discussion

Table 1 shows the segmentation results of methods on the SemanticKITTI [1] validation set. Compared to image-based methods, the proposed approach attains superior segmentation performance for all three evaluated input sizes, namely 64×512, 64×1024, and 64×2048. For the distance image with an input size of 64 × 2048, the proposed method achieves state-of-the-art performance (69.5% mIoU), outperforming point-based methods, voxel-based methods, and other compared approaches. The best segmentation performance among point-based methods reaches only 59.9%, which is much lower than the proposed method. With an input size of 64 × 1024, the proposed method achieves 68.9% mIoU, representing a 3.1% improvement over the benchmark model, CENet. For an input size of 64 × 512, the proposed method achieves 67.5% mIoU, a maximum improvement of 4.3% over previous methods. These results highlight the accuracy and robustness of the proposed method under different input resolutions.

Additionally, the proposed method significantly improves segmentation performance in 15 out of the 19 categories of the SemanticKITTI dataset. Notably, for dynamic objects (e.g., pedestrians), the mIoU improves by up to 10% at different scales. It is worth noting that these performance improvements are achieved with only a 0.7% increase in the number of model parameters. The impact of this parameter increase will be analyzed in detail in the subsequent ablation experiments, focusing on its effect on model performance and efficiency.

Table 2 compares the performance of the proposed method with other related works on the SemanticPOSS [2] test set. Specifically, the proposed method shows significant improvements over the baseline model in terms of mIoU: with KNN, it outperforms CENet [12] by 2.6% mIoU, FIDNet [11]by 6.7% mIoU, and MINet [37] by 9.8% mIoU; without KNN, it outperforms CENet [12] by 2.4% mIoU, FIDNet [11] by 6.3% mIoU, and MINet [37] by 9.5% mIoU. These results further demonstrate the strong generalization ability and robustness of GASegNet, the network proposed in this paper. GASegNet achieves an mIoU score of 52.7%, outperforming most categories in terms of IoU metrics across the board.

C. Qualitative Results

Qualitative Result On SemanticKITTI: For further visualization, we provide examples of qualitative comparisons in Fig. 5. As shown in Fig. 5a, Fig. 5b, and Fig. 5f, the proposed method recognizes traffic signs, motorcycles, and vegetation more accurately than CENet [12], even as complete objects. This improvement is attributed to the feature fusion module, which enhances the characterization of local features and strengthens the local perception capability of the segmentation network model. This is particularly evident in Fig. 5c, where CENet [12] misidentifies the upper part of a person as a motorcycle and the lower part as a person, a common issue in point cloud semantic segmentation networks. In contrast, the proposed method accurately recognizes the entire person.

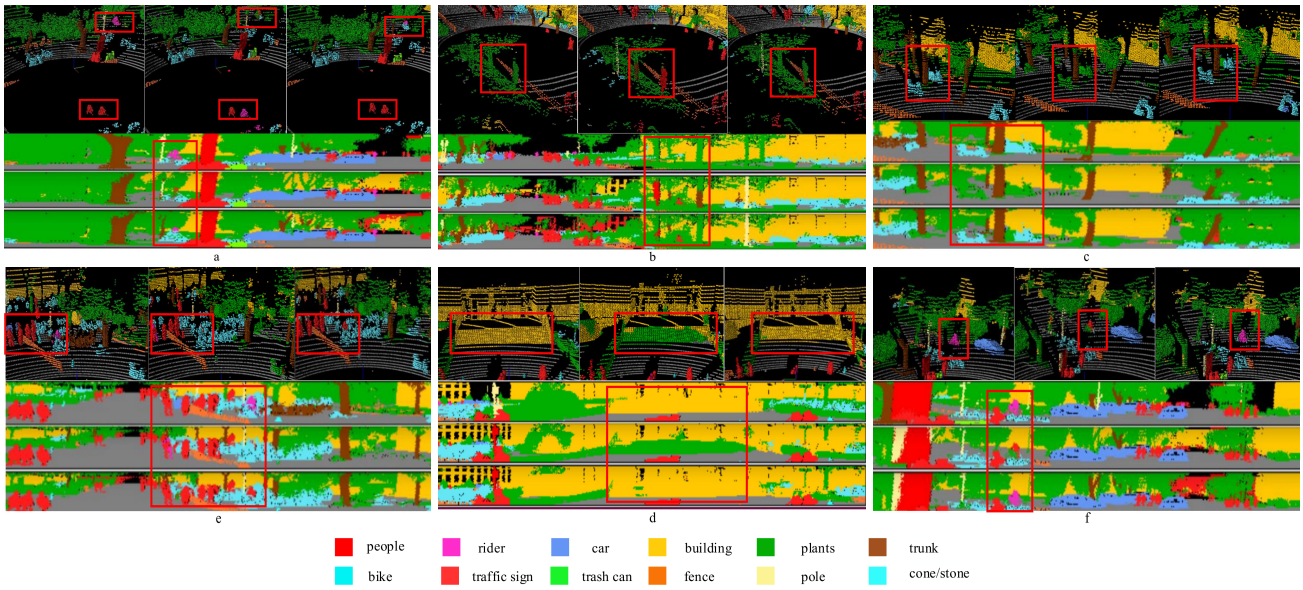

Fig. 6. Qualitative analysis of the SemanticPOSS validation set. The 3D point cloud at the top of the image presents, from left to right, the real data, the semantic labels predicted by the CENet model, and those predicted by our model, GASegNet. Below the 3D point cloud is a depth map with semantic labels displayed in rows, showing, from top to bottom, the real values, CENet predictions, and GASegNet predictions.

In Figs. 5d and 5e, the labeled person is partially occluded. The proposed method successfully identifies the occluded category, demonstrating its ability to address the point cloud occlusion problem. This is due to the global attention mechanism, which effectively constructs long-distance dependencies of spatial locations and captures global context information. In Fig. 5f, the baseline model CENet [12] fails to segment vegetation and fences correctly, while the proposed method demonstrates improved performance.

Qualitative Result On SemanticPOSS: Figure 6 presents qualitative analysis on the SemanticPOSS validation set. In Fig. 6a, Fig. 6e, and Fig. 6f, our model accurately recognizes dynamic objects like riders and people. In contrast, CENet often misclassifies riders as people, while GASegNet distinguishes them effectively. As shown in Fig. 6c, CENet [12] misidentifies bicycles parked next to tree trunks as plants due to occlusion, whereas our method leverages global context information to mitigate this. Additionally, Fig. 6b and Fig. 6d show improved recognition of static objects.

The quantitative and qualitative analyses confirm the effectiveness of our point cloud semantic segmentation network, which integrates a global attention mechanism and a feature fusion module. The proposed method excels in handling both static and dynamic objects in complex urban environments. By effectively aggregating local and global features, the model addresses rapid object motion and occlusion in complex dynamic scenes. This enables robust and accurate feature extraction, making the model highly suitable for autonomous driving and robotic navigation in real-world environments.

D. Ablation Experiments

To quantitatively evaluate the contribution of each component, we conduct ablation experiments on Sequence 08 of the SemanticKITTI validation set. Following CENet's training and evaluation settings (input resolution: 64×512), we ensure fair comparison. Table 3 reports mIoU and parameter counts, highlighting the impact of each module on segmentation accuracy and efficiency.

From Table 3, the integration of FFM improves the mIoU by 1.7 points, from 63.2 to 64.9, demonstrating its effectiveness in enhancing segmentation performance. This is attributed to FFM's ability to reorganize geometric features, strengthening local feature learning and scene understanding. Additionally, the embedded attention mechanism addresses the uneven distribution of point cloud data, optimizing feature separation. As a result, the model achieves better segmentation accuracy while maintaining computational efficiency.

With the introduction of GSAM, the model's mIoU increases to 67.0, representing a 3.8% improvement over CENet, achieved with only a parameter increase (0.02M). This significant gain is due to GSAM's ability to capture global contextual features and strengthen the connection between local and global information via skip connections. By leveraging global attention, GSAM is capable of capturing long-range dependencies within the point cloud and effectively addresses occlusion and model complexity, enhancing scene understanding while maintaining lightweight computational efficiency.

Our proposed GASegNet method shows a 4.3% improvement compared to the benchmark model CENet, with only a 0.7% increase in parameters. Seen, compared to the benchmark model CENet, our method demonstrates an improvement in mIoU across all experiments. This validates the effectiveness of each module and shows that our improved method significantly enhances the performance of semantic segmentation of point clouds while keeping the model lightweight.

TABLE I
EVALUATION RESULT ON THE SEMANTICKITTI VALIDATION SET

Category	Methods	Size	**mIoU**	Car	bicycle	motorcycle	truck	Other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
Point-based	PointNet++ [4]	50Kpts	20.1	53.7	1.9	0.2	0.9	0.2	0.9	1.0	0	72.0	18.7	41.8	5.6	62.3	16.9	46.5	13.8	30.0	6.0	8.9
	RandLa-Net [10]		53.9	94.2	26.0	25.8	40.1	38.9	49.2	48.2	7.2	90.7	60.3	73.7	20.4	86.9	56.3	81.4	61.3	66.8	49.2	47.7
	KPConv [13]		58.8	**96.0**	30.2	**42.5**	33.4	44.3	**61.5**	**61.6**	11.8	88.8	61.3	72.7	**31.6**	**90.5**	**64.2**	**84.8**	**69.2**	**69.1**	**56.4**	47.4
	BAAF [14]		**59.9**	95.4	**31.8**	35.5	**48.7**	**46.7**	49.5	55.7	**53.0**	**90.9**	**62.2**	**74.4**	23.6	89.8	60.8	82.7	63.4	67.9	53.7	**52.0**
Voxel-based	MinkowskiNet-lite [38]	Voxel	57.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	MinkowskiNe [38]		63.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	SPVCNN-lite [39]		58.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	SPVCNN [39]		**63.8**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Image-based	SqueezeSeg [22]	64 x 2048	30.8	68.3	18.1	5.1	4.1	4.8	16.5	17.3	1.2	84.9	28.4	54.7	4.6	61.5	29.2	59.6	25.5	54.7	11.2	36.3
	SqueezeSegV2 [23]		39.6	82.7	21.0	22.6	14.5	15.9	20.2	24.3	2.9	88.5	42.4	65.5	18.7	73.8	41.0	68.5	36.9	58.9	12.9	41.0
	SqueezeSegV3 [9]		55.9	92.5	38.7	36.5	29.6	33.0	45.6	46.2	20.1	91.7	63.4	74.8	26.4	89.0	59.4	82.0	58.7	65.4	49.6	58.9
	SalsaNext [24]		59.5	91.9	**48.3**	38.6	**38.9**	31.9	60.2	59.0	19.4	91.7	63.7	75.8	29.1	90.2	64.2	81.8	63.6	66.5	54.3	62.1
	Lite-HDSeg [27]		**63.8**	92.3	40.0	**55.4**	37.7	39.6	59.2	**71.6**	**54.3**	93.0	68.2	78.3	29.3	91.5	65.0	78.2	65.8	65.1	**59.5**	**67.7**
	KPRNet [26]		63.1	**95.5**	54.1	47.9	23.6	**42.6**	**65.9**	65.0	16.5	**93.2**	**73.9**	**80.6**	**30.2**	**91.7**	**68.4**	**85.7**	**69.8**	**71.2**	58.7	64.1
	RangeNet++ [8]		41.9	87.4	26.2	26.5	18.6	15.6	31.8	33.6	4.0	91.4	57.0	74.0	26.4	81.9	52.3	77.6	48.4	63.6	36.0	50.0
	MPF [40]	64 x 512	48.9	91.1	22.0	19.7	18.8	16.5	30.0	36.2	4.2	91.1	**61.9**	74.1	**29.4**	86.7	56.2	82.3	51.6	68.9	38.6	49.8
	FIDNet [11]		51.3	90.4	28.6	30.9	34.3	27.0	43.9	48.9	**16.8**	90.1	58.7	71.4	19.9	84.2	51.2	78.2	51.9	64.5	32.7	50.3
	CENet [12]		63.2	96.1	47.4	55.4	**79.2**	54.9	60.9	72.5	0.1	95.7	60.4	83.1	4.1	86.0	59.8	87.2	65.9	73.6	68.4	49.2
	Ours		**67.5**	**96.4**	**55.7**	**70.5**	75.9	**64.3**	**73.4**	**86.2**	1.3	**95.9**	58.8	**84.5**	9.2	**87.1**	**61.4**	**88.8**	**68.5**	**78.1**	**70.2**	**56.0**
Image-based	RangeNet++ [8]	64 x 1024	48.0	90.3	20.6	27.1	25.2	17.6	29.6	34.2	7.1	90.4	52.3	72.7	22.8	83.9	53.3	77.7	52.5	63.7	43.8	47.2
	MPF [40]		53.6	92.7	28.2	30.5	26.9	25.2	42.5	45.5	9.5	90.5	**64.7**	74.3	**32.0**	88.3	59.0	83.4	56.6	69.8	46.0	54.9
	FIDNet [11]		56.0	92.4	44.0	41.5	33.2	30.8	57.9	52.6	**18.0**	91.0	61.2	73.8	12.6	88.2	57.9	80.8	59.5	65.1	45.3	58.4
	CENet [12]		65.8	**96.8**	**60.1**	70.7	67.0	**66.5**	77.3	87.7	0.0	95.7	44.8	83.2	12.8	83.9	46.0	86.9	69.9	72.3	**70.4**	57.6
	Ours		**68.9**	96.3	56.5	**75.9**	**83.5**	57.8	**80.2**	**88.5**	0.0	**96.0**	55.1	**84.5**	9.7	**88.9**	**64.5**	**89.7**	**71.1**	**79.8**	69.7	**61.5**
Image-based	RangeNet++ [8]	64 x 2048	52.2	91.4	25.7	34.4	25.7	23.0	38.3	38.8	4.8	91.8	**65.0**	75.2	27.8	87.4	58.6	80.5	55.1	64.6	47.9	55.9
	MPF [40]		55.5	93.4	30.2	38.3	26.1	28.5	48.1	46.1	18.1	90.6	62.3	74.5	**30.6**	88.5	59.7	83.5	59.7	69.2	49.7	58.1
	FIDNet [11]		58.6	93.0	45.7	42.0	27.9	32.6	62.6	58.1	**30.5**	90.8	58.3	74.9	20.1	88.5	59.5	83.1	64.3	67.8	52.6	60.0
	CENet [12]		65.5	96.1	55.2	68.72	87.2	48.0	72.6	74.2	0.0	95.2	52.6	**83.2**	11.4	88.79	50.4	87.9	69.3	**76.5**	65.3	61.1
	Ours		**69.5**	**97.5**	**62.2**	**78.8**	**89.1**	**73.5**	**82.8**	**88.8**	0.0	**95.9**	32.9	83.0	0.1	**91.3**	**71.0**	**88.3**	**73.3**	76.4	**69.5**	**66.4**

TABLE II
EVALUATION RESULTS ON THE SEMANTICPOSS TEST SPLIT

Methods	person	rider	car	truck	plants	traffic-sign	pole	trashcan	building	cone/stone	fence	bike	ground	mIoU
SqueezeSeg [22]	14.9	1.0	13.2	10.4	28.0	5.1	5.7	2.3	43.6	0.2	15.6	31.0	75.0	18.9
SqueezeSeg + CRF [22]	6.8	0.6	6.7	4.0	2.5	9.1	1.3	0.4	37.1	0.2	8.4	18.5	72.1	12.9
SqueezeSegV2 [23]	48.0	9.4	48.5	11.3	50.1	6.7	6.2	14.8	60.4	5.2	22.1	36.1	71.3	30.0
SqueezeSegV2 + CRF [23]	43.9	7.1	47.9	18.04	40.9	4.8	2.8	7.4	57.5	0.6	12.0	35.3	71.3	26.9
RangeNet53 [8]	55.7	4.5	34.4	13.7	57.5	3.7	6.6	23.3	64.9	6.1	22.2	28.3	72.9	30.3
RangeNet53 + KNN [8]	57.3	4.6	35.0	14.1	58.3	3.9	6.9	24.1	66.1	6.6	23.4	28.6	73.5	30.9
MINet [37]	61.8	12.0	63.3	22.2	68.1	16.3	29.3	28.5	74.6	25.9	31.7	44.5	76.4	42.7
MINet + KNN [37]	62.4	12.1	63.8	22.3	68.6	16.7	30.1	28.9	75.1	28.6	32.2	44.9	76.3	43.2
FIDNet-Point [11]	71.6	22.7	71.7	22.9	67.7	21.8	27.5	15.8	72.7	31.3	40.4	50.3	79.5	45.8
FIDNet-Point + KNN [11]	72.2	23.1	72.7	23.0	68.0	22.2	28.6	16.3	73.1	34.0	40.9	50.3	79.1	46.4
CENet [12]	74.9	21.8	77.0	25.3	72.0	18.0	30.9	46.9	75.9	26.1	47.5	51.7	**80.7**	49.9
CENet + KNN [12]	**75.5**	22.0	77.6	25.3	72.2	18.2	31.5	48.1	76.3	27.7	47.7	51.4	80.3	50.3
Ours	75.3	23.8	77.5	27.2	73.3	27.9	32.0	50.3	78.6	33.8	**48.0**	**54.1**	80.5	52.5
Ours + KNN	75.2	**24.4**	**78.2**	**27.6**	**73.8**	**28.2**	**33.0**	**52.4**	**79.2**	**34.9**	46.6	53.1	78.6	**52.7**

TABLE III
ABLATION EXPERIMENTS EVALUATED ON SEMANTICKITTI

Baseline	FFM	GSAM	mIoU	Params (M)
CENet [12]			63.2	6.783
Ours	✓		64.9	6.813
Ours		✓	67.0	6.803
Ours	✓	✓	67.5	6.833

V. CONCLUSION

In this paper, we propose GASegNet, a novel semantic segmentation framework tailored for LiDAR point cloud tasks. By integrating a feature fusion module and a global self-attention mechanism within an encoder-decoder architecture, GASegNet enhances feature representation. Additionally, the incorporation of boundary loss emphasizes semantic boundaries, while auxiliary segmentation heads further boost feature learning capability without increasing parameter count or computational cost. The performance of GASegNet was evaluated

on the SemanticKITTI and SemanticPOSS datasets, where it achieved top-tier results in semantic segmentation and real-time processing, confirming its high capability. Despite these advances, there is still room for improvement, particularly in recognizing cyclists and certain static objects. Constructing better local feature representations can enhance the network's ability to extract and learn local features. In future work, our team will focus on advancing semantic segmentation through multimodal data fusion.

REFERENCES

[1] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9297–9307, doi: 10.1109/ICCV.2019.00939.

[2] Y. Pan, B. Gao, J. Mei, S. Geng, C. Li, and H. Zhao, "Semanticposs: A point cloud dataset with large quantity of dynamic instances," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 687–693, doi: 10.1109/IV47402.2020.9304596.

[3] C. R. Qi, H. Su, K. Mo, and 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*, 2017, pp. 652–660, doi: 10.1109/CVPR.2017.16.

[4] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017, doi: 10.48550/arXiv.1706.02413.

[5] X. Wu, L. Jiang, P.-S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, and H. Zhao, "Point transformer v3: Simpler faster stronger," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 4840–4851, doi: 10.1109/CVPR52733.2024.00463.

[6] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017, doi: 10.48550/arXiv.1710.10903.

[7] X. Yan, C. Zheng, Z. Li, S. Wang, and S. Cui, "Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5588–5597, doi: 10.1109/CVPR42600.2020.00563.

[8] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "Rangenet++: Fast and accurate lidar semantic segmentation," in *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2019, pp. 4213–4220, doi: 10.1109/IROS40897.2019.8967762.

[9] C. Xu, B. Wu, Z. Wang, W. Zhan, P. Vajda, K. Keutzer, and M. Tomizuka, "Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*. Springer, 2020, pp. 1–19, doi: 10.1007/978-3-030-58604.

[10] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "Randla-net: Efficient semantic segmentation of large-scale point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 108–11 117, doi: 10.1109/CVPR42600.2020.01112.

[11] Y. Zhao, L. Bai, and X. Huang, "Fidnet: Lidar point cloud semantic segmentation with fully interpolation decoding," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4453–4458, doi: 10.1109/IROS51168.2021.9636385.

[12] H.-X. Cheng, X.-F. Han, and G.-Q. Xiao, "Cenet: Toward concise and efficient lidar semantic segmentation for autonomous driving," in *2022 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2022, pp. 01–06, doi: 10.1109/ICME52920.2022.9859693.

[13] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6411–6420, doi: 10.1109/ICCV.2019.00651.

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

[15] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019, doi: 10.1145/3326362.

[16] C. Chen, L. Z. Fragonara, and A. Tsourdos, "Gapointnet: Graph attention based point neural network for exploiting local feature of point cloud," *Neurocomputing*, vol. 438, pp. 122–132, 2021, doi: 10.1016/j.neucom.2021.01.095.

[17] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 922–928, doi: 10.1109/IROS.2015.7353481.

[18] R. Klokov and V. Lempitsky, "Escape from cells: Deep kd-networks for the recognition of 3d point cloud models," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 863–872, doi: 10.1109/ICCV.2017.99.

[19] H. Su, V. Jampani, D. Sun, S. Maji, E. Kalogerakis, M.-H. Yang, and J. Kautz, "Splatnet: Sparse lattice networks for point cloud processing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2530–2539, doi: 10.1109/CVPR.2018.00268.

[20] X. Zhu, H. Zhou, T. Wang, F. Hong, Y. Ma, W. Li, H. Li, and D. Lin, "Cylindrical and asymmetrical 3d convolution networks for lidar segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9939–9948, doi: 10.1109/TPAMI.2021.3098789.

[21] R. Cheng, R. Razani, E. Taghavi, E. Li, and B. Liu, "2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 547–12 556, doi: 10.1109/CVPR46437.2021.01236.

[22] B. Wu, A. Wan, X. Yue, and K. Keutzer, "Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1887–1893, doi: 10.1109/ICRA.2018.8462926.

[23] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 4376–4382, doi: 10.1109/ICRA.2019.8793495.

[24] T. Cortinhal, G. Tzelepis, and E. Erdal Aksoy, "Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds," in *Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II 15*. Springer, 2020, pp. 207–222, doi: 10.1007/978-3-030-64559.

[25] E. E. Aksoy, S. Baci, and S. Cavdar, "Salsanet: Fast road and vehicle segmentation in lidar point clouds for autonomous driving," in *2020 IEEE intelligent vehicles symposium (IV)*. IEEE, 2020, pp. 926–932, doi: 10.1109/IV47402.2020.9304694.

[26] D. Kochanov, F. K. Nejadasl, and O. Booij, "Kprnet: Improving projection-based lidar semantic segmentation," *arXiv preprint arXiv:2007.12668*, 2020, doi: 10.48550/arXiv.2007.12668.

[27] R. Razani, R. Cheng, E. Taghavi, and L. Bingbing, "Lite-hdseg: Lidar semantic segmentation using lite harmonic dense convolutions," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 9550–9556, doi: 10.1109/ICRA48506.2021.9561171.

[28] J. Li, Y. Wen, and L. He, "Scconv: Spatial and channel reconstruction convolution for feature redundancy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6153–6162, doi: 10.1109/CVPR52729.2023.00596.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012, doi: 10.1145/3065386.

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9, doi: 10.1109/cvpr.2015.7298594.

[31] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017, doi: 10.48550/arXiv.1704.04861.

[32] S. Wang, J. Zhu, and R. Zhang, "Meta-rangeseg: Lidar sequence semantic segmentation using multiple feature aggregation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9739–9746, 2022, doi: 10.1109/LRA.2022.3191040.

[33] A. Athar, E. Li, S. Casas, and R. Urtasun, "4d-former: Multimodal 4d panoptic segmentation," in *Conference on Robot Learning*. PMLR, 2023, pp. 2151–2164, doi: 10.48550/ARXIV.2311.01520.
[34] D. Ye, W. Chen, Z. Zhou, Y. Xie, Y. Wang, P. Wang, and H. Foroosh, "Lidarmultinet: Unifying lidar semantic segmentation, 3d object detection, and panoptic segmentation in a single multi-task network," *arXiv preprint arXiv:2206.11428*, 2022, doi: 10.48550/arXiv.2209.09385.
[35] Y. Liu, R. Chen, X. Li, L. Kong, Y. Yang, Z. Xia, Y. Bai, X. Zhu, Y. Ma, Y. Li *et al.*, "Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 662–21 673, doi: 10.1109/ICCV51070.2023.01980.
[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017, doi: 10.48550/arXiv.1711.05101.
[37] S. Li, X. Chen, Y. Liu, D. Dai, C. Stachniss, and J. Gall, "Multi-scale interaction for real-time lidar data segmentation on an embedded platform," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 738–745, 2021, doi: 10.1109/LRA.2021.3132059.
[38] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9224–9232, doi: 10.1109/CVPR.2018.00961.
[39] H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching efficient 3d architectures with sparse point-voxel convolution," in *European conference on computer vision*. Springer, 2020, pp. 685–702, doi: 10.1007/978-3-030-58604.
[40] Y. A. Alnaggar, M. Afifi, K. Amer, and M. ElHelw, "Multi projection fusion for real-time semantic segmentation of 3d lidar point clouds," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1800–1809, doi: 10.1109/WACV48630.2021.00184.

Xu Lu received the B.S. degrees from Nanchang University, Jiangxi, China, in 2006, and the M.E. and Ph.D. degree from the Guangdong University of Technology, Guangdong, China, in 2009 and 2015, respectively. His research interests include artificial intelligence and smart system.

Haijun Liu is currently pursuing a master's degree in artificial intelligence at the Institute of Interdisciplinary Studies, Guangdong Polytechnic Normal University. His main research directions include robotics and artificial intelligence.

Guang'an Luo is currently pursuing a master's degree in New-Generation Electronic Information Technology at the School of Electronics and Information Technology, Guangdong Polytechnic Normal University. His main research directions include the Internet of Things and artificial intelligence.

Zhike Chen is currently pursuing a master's degree in Control Science and Engineering at the School of Computer Science, Guangdong Polytechnic Normal University. His main research directions include the Internet of Things and machine vision.

Cheng Zhou is currently pursuing a master's degree in Control Science and Engineering at the School of Computer Science, Guangdong Polytechnic Normal University. His main research directions include the Internet of Things and artificial intelligence.

Xinyu Wu (Senior Member, IEEE) received the B.E. and M.E.degrees from the Department of Automation, University of Science and Technology of China, Hefei,China, in 2001 and 2004, respectively. His Ph.D. degreen was awarded at the Chinese University of Hong Kong in 2008. He is currently a Professor with Shenzhen Institute of Advanced Technology, Shenzhen,China, the Director of Center for Intelligent Bionic, and the Director of the Guangdong Provincial Key Lab of Robotics and Intelligent System. He has authored or co-authored more than 260 journal and conference papers and 2 monographs. His research interests include wearable robotics, human-machine interaction, and intelligent system. He received the GaiTech Best Paper in Robotics Award at the IEEE International Conference on Information and Automation (ICIA) in 2018 and the Best Application Paper Award at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) in 2019, among others. Professor Wu has been an Associate Editor of several journals, including IEEE Transactions on Systems Man Cybernetics-Systems, IEEE Transactions on Automation Science and Engineering, and IEEE Robotics and Automation Letters.

Jun Liu received the B.S. degree in electronic information engineering in 2009 from Qiqihar University, Qiqihar, China, and the M.S. and Ph.D. degrees in control science and engineering in 2012 and 2015, respectively, from Guangdong University of Technology, Guangzhou, China. He is currently an associate professor at Guangdong Polytechnic Normal University in Guangzhou, China. His research interests mainly include VSLAM and intelligent mobile.