SGFormer: Semantic-Geometry Fusion Transformer for Multi-modal 3D Panoptic Segmentation

Hongqi Yu[1], **Sixian Chan**[2*], **Xiaolong Zhou**[3], **Xiaoqin Zhang**[1*]

[1]Key Laboratory of Intelligent Informatics for Safety and Emergency of Zhejiang Province, Wenzhou University, China
[2]College of Computer Science and Technology, Zhejiang University of Technology, China
[3]College of Electrical and Information Engineering, Quzhou University, China
yuhongqcn@163.com, sxchan@zjut.edu.cn, xiaolong@ieee.org, zhangxiaoqinnan@gmail.com

Abstract

Modern methods for autonomous driving perception widely adopt multi-modal fusion to enhance 3D scene understanding. However, existing methods suffer from inferior semantic extraction in image encoders that treat all pixels equally, ignoring contextual differences. The generated multi-modal representations also typically lack comprehensive semantic and spatial geometry information, which is crucial for the 3D panoptic segmentation task. In this paper, we propose a novel **S**emantic-**G**eometry **F**usion Trans**former** (**SGFormer**) that extracts adaptive semantic contexts, aggregates geometric information and captures the semantic-geometry fusion. First, in the Image Branch, we tailor semantic contexts for each pixel with context-guided attention and spatial context alignment to refine semantic details. Second, we transform image and voxel features into point-pixel geometry representations, simultaneously learning semantic category priors as embeddings to better represent scene geometry and semantics. Finally, to aggregate semantic information with related geometry, we design a semantic-geometry fusion that combines the transformer, effectively capturing semantic-geometry relationships into multi-modal panoptic representations. Notably, SGFormer achieves the state-of-the-art (SOTA) results on the nuScenes and SemanticPOSS, as well as yielding competitive performance on the SemanticKITTI. Moreover, SGFormer exhibits superior robustness compared to leading methods, marking an improvement of 2% to 10%.

Introduction

3D scene understanding has been a critical task for autonomous driving. One of the key tasks is 3D panoptic segmentation, comprising semantic and instance segmentation as two sub-tasks. The semantic segmentation aims to assign semantic labels at the point level, while the instance segmentation intends to identify individual countable objects.

Point clouds and images are two complementary modalities in perception tasks. Point clouds, while rich in spatial information, are sparse and often struggle to distinguish between foreground and background. While images can provide rich visual information on color and shape to generate semantic contexts. Hence, a recent trend has emerged to exploit semantic features from image segmentation to improve

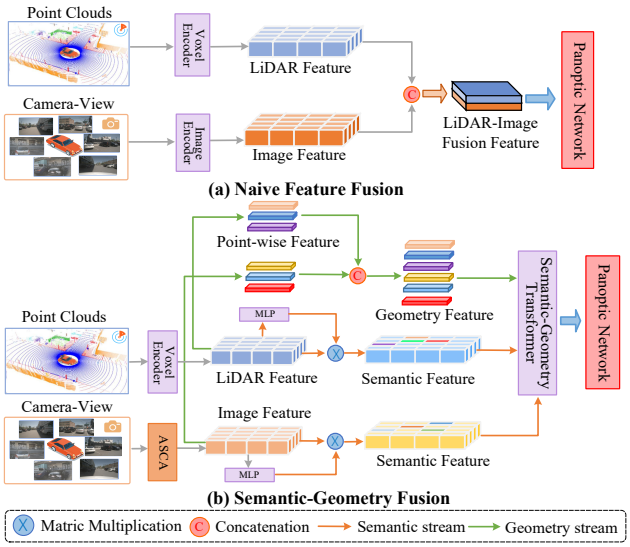


Figure 1: Comparison (a) naive fusion using feature concatenation (b) our semantic-geometry fusion integrating semantic and geometric information.

perception. Image semantic segmentation aims to assign a predefined category label to each pixel in the image. Many works adopt deep networks to explore scene contexts (Yang et al. 2018) and spatial details (Lin et al. 2017a) to ensure accurate semantics. However, they inevitably suffer from high computation. To achieve real-time inference, various efforts are made by designing lightweight networks (Lv et al. 2021) or efficient decoders (Peng et al. 2022). Despite reducing computation, information loss in lightweight networks leads to misalignment of input and output. Besides, context modeling methods (Dong et al. 2020; Xu, Xiong, and Bhattacharyya 2023) in existing decoders also lack adaptability to different inputs. Moreover, we notice most approaches (Liu et al. 2023; Zhang, Zhu, and Du 2023) in 3D tasks (*e.g.,* detection, completion) adopt image backbones with FPN (Lin et al. 2017b) to integrate multi-scale features via stepwise downsampling, which treats all pixels equally without contextual differences and overlooks representation differences between different level features. These motivate us to consider: *can we aggregate adaptive pixel contexts from*

*Corresponding authors

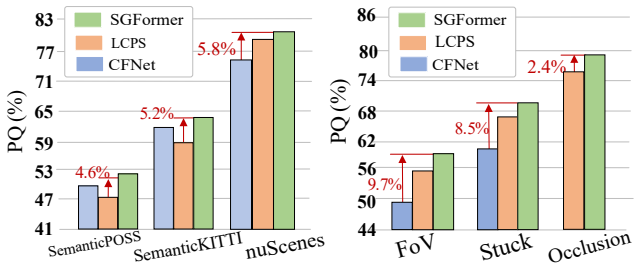


Figure 2: 3D panoptic results of different methods on performance and robustness. Our proposed method achieves 2%-10% relative improvement.

image-modal to provide accurate semantic information?

Further revisiting the difference between LiDAR and image modalities that have been widely used in 3D vision (Yan et al. 2023; Pan, Wang, and Wang 2024), we observed that it is beneficial to utilize multi-modal information for enhancing 3D perception. Yet, existing 3D panoptic segmentation methods are primarily based on LiDAR data alone (Li et al. 2023b; Xiao et al. 2023), which means that efficient image and LiDAR fusion for capturing multi-modal information remains a challenging problem. LCPS (Zhang et al. 2023), a recent LiDAR-camera panoptic network, has demonstrated the potential of multi-modal fusion in achieving 3D panoptic segmentation. However, the generated multi-modal representations still lack comprehensive semantic and geometry information. Meanwhile, considering the differences between LiDAR-modal with positional geometry and image-modal with semantic context. We further ponder: *how to efficiently capture such semantic-geometry fusion to generate fine-grained multi-modal panoptic representations?*

With the above observations, we introduce a **S**emantic-**G**eometry **F**usion Trans**former** for 3D panoptic segmentation, termed **SGFormer**. As shown in Fig. 1, our SGFormer achieves semantic-geometry fusion within a transformer structure through aggregating semantic- and geometry-wise information. Specifically, we devise a new image branch with Adaptive Semantic-Context Aggregation (ASCA) for aggregating adaptive contexts to pixels, which includes Context-Guided Attention (CGA), Spatial Context Alignment (SCA) and Point-Pixel Refinement (PPR). Firstly, the CGA extracts semantic contexts for pixels with pixel-context attention to enhance semantic discrimination. Then, the SCA is designed to fuse cross-level features adaptively with spatial detail alignment. Next, the PPR aims to establish point-to-pixel mapping for pixel feature refinement.

Further, to generate fine-grained panoptic representations enriched with semantic and geometry, we also propose a novel structure called SGTransformer that adopts a transformer sub-network to model the complex relationships of semantic context and spatial geometry. The SGTransformer involves three phases: i) acquiring 3D voxel point-wise and 2D image point-wise features to generate point-pixel geometry representations; ii) utilizing MLP-based semantic classifiers to learn semantic category priors from both modalities as feature embeddings; iii) employing a semantic-geometry fusion module to model potential multi-modal semantic interactions based on self-attention. Meanwhile, to facilitate fusion of semantic- and geometry-wise features, it uses cross-attention to aggregate more cross-modal semantic information related geometric points. The proposed SGFormer effectively captures model-related semantic contexts and spatial geometry, surpassing existing methods with a relative improvement of 2%-10%, as illustrated in Fig. 2.

In summary, our contributions are summarized as follows:

- We introduce SGFormer, a novel panoptic segmentation method based on semantic-geometry multi-modal fusion.
- We design a simple yet effective adaptive semantic-context aggregation to extract adaptive semantic contexts for pixels, providing accurate semantic features.
- We propose a semantic-geometry transformer to aggregate semantic information and spatial geometry, including a semantic-geometry fusion module to capture semantic-geometry relationships into fine-grained multi-modal panoptic representations.
- SGFormer achieves state-of-the-art performance on 3D panoptic segmentation task. Moreover, it exhibits superior robustness compared to leading methods.

Related Work

Vision-Based Semantic Segmentation. Image Semantic segmentation is extensively applied in automatic driving. Previous methods for semantic segmentation are based on FCNs (Long, Shelhamer, and Darrell 2015), while they have an inherent weakness, *i.e.,* the local receptive fields. Recent progresses mainly employ two strategies to improve performance, *i.e.,* context modeling (Zhu et al. 2019) and feature fusion (Huang et al. 2021). But context modeling in existing decoders lacks adaptability to different inputs and feature fusion often ignores feature misalignment problem. Thus, some works (Fu et al. 2019; Yuan et al. 2021) exploit self-attention to encode the global contexts for each pixel adaptively, but they suffer from high computation. Compared to these works, we investigate an adaptive semantic context aggregation from a new perspective, leveraging pix-context attention and multi-level spatial alignment, and thus providing accurate image-modal semantic information.

LiDAR-Based Panoptic Segmentation. The panoptic segmentation task is first proposed in the image domain (Kirillov et al. 2019) and later extended to LiDAR point cloud dataset, *e.g.,* SemanticKITTI (Behley et al. 2019) and nuScenes (Fong et al. 2022). The LiDAR-based 3D panoptic segmentation can be classified into proposal-based and proposal-free. Proposal-based methods (Sirohi et al. 2021; Ye et al. 2023) firstly locate objects, predict instance mask for each bounding box, and then fuse the two results. This paradigm makes the performance inevitably affected by object detection. While proposal-free methods (Hong et al. 2021; Li et al. 2022, 2023a) firstly conduct semantic segmentation and then cluster the "thing" points belonging to instances based on object center and point offset.

Multi-modal Panoptic Segmentation. Images and LiDAR points have made considerable progress in 3D perception

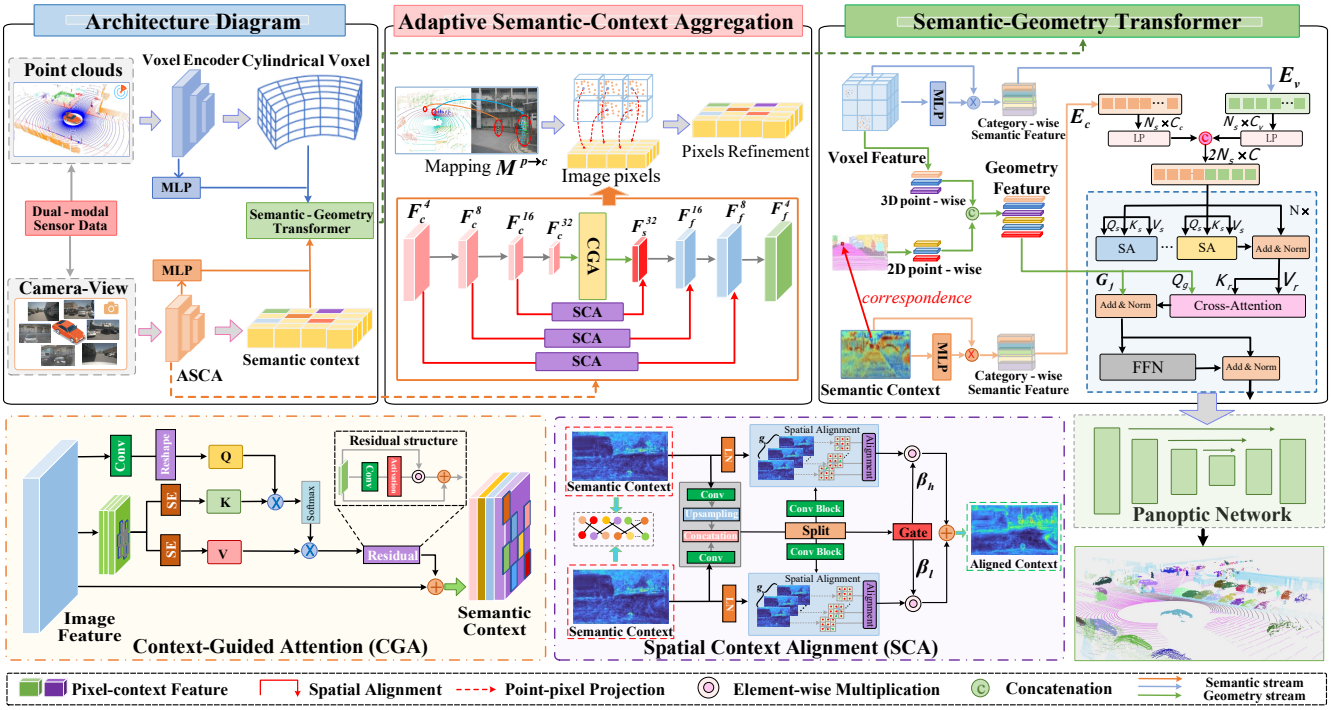


Figure 3: Workflow of the proposed SGFormer, comprising two key modules. The Adaptive Semantic-Context Aggregation extracts semantic-dependent contexts to pixels with context-guided attention and spatial context alignment. The Semantic-Geometry Transformer generates fine-grained multi-modal panoptic representations through semantic-geometry fusion.

tasks, *e.g.,* object detection (Bai et al. 2022; Liu et al. 2023) and semantic occupancy (Pan, Wang, and Wang 2024; Zhang and Ding 2024). However, multi-modal methods for 3D panoptic segmentation remain underexplored. Unlike occupancy tasks, which assign semantic labels based on the approximate location of voxels, panoptic segmentation requires accurate spatial position and semantic information to label each point. Recently, LCPS (Zhang et al. 2023) considers LiDAR-camera fusion for enhancing perception, bur it is limited to point-pixel consistent alignment for modal-fusion, resulting in unsatisfactory performance on SemanticKITTI when fewer camera images available. In contrast, we adopt a semantic-geometry fusion for capturing multi-modal information by aggregating point-wise geometry features and learning semantic category priors as feature embeddings.

Methodology

Preliminary

Given input point clouds $\mathbf{L} = \{p_i | p_i \in \mathbb{R}^{3+c}\}_{i=1}^{N_p}$ of N_p points, each point has $(3+c)$-dimensional features and camera images $\mathbf{I} = \{I_i | I_i \in \mathbb{R}^{3\times H\times W}\}_{i=1}^{N_c}$ from N_c surround-view cameras. 3D panoptic segmentation aims to precisely predict unique semantic labels $\mathbf{S} = \{s_i | s_i \in (1 \cdots K)\}_{i=1}^{N_p}$ for the points, where K is the number of predicted semantic classes, and identify points belonging to separate instances with unique instance ID in some thing classes, *e.g.,* car, bicycle, and pedestrian. In our work, the panoptic segmentation is formulated as an essential semantic-geometry aggregation and semantic-geometry fusion problem.

Overall Architecture

An overview of Semantic-Geometry Fusion Transformer (SGFormer) is presented in Fig. 3. The SGFormer mainly consists of two key modules: an adaptive semantic-context aggregation to extract adaptive contexts to pixels for providing semantic features and a semantic-geometry transformer structure to generate multi-modal panoptic representations enriched with semantic and geometry information.

Following voxel encoder (*e.g.,* Cylinder Convolution (Zhu et al. 2021)) and image encoder (*e.g.,* ResNet-18 (He et al. 2016)) for LiDAR and image extraction, we first take raw points $\mathbf{L}$ of N_p points and images $\mathbf{I}$ of N_c cameras as input to obtain voxel-modal features $\mathbf{F}_\mathrm{v} \in \mathbb{R}^{N_v\times C_v}$ and multi-scale image-modal features $\mathbf{F}_\mathrm{c}^{32}$, $\mathbf{F}_\mathrm{c}^{16}$, $\mathbf{F}_\mathrm{c}^{8}$, $\mathbf{F}_\mathrm{c}^{4} \in \mathbb{R}^{N_c\times C_c\times H_c\times W_c}$, where N_v, C_v are the number of voxels and feature dimensions, and C_c, H_c, W_c represent the dimensions and size of image features. In the following sections, we provide detailed descriptions of each module.

Adaptive Semantic-Context Aggregation

Multi-modal information has gained attention in recent 3D tasks. However, the 2D image branch has been somewhat neglected, leaving the network far from better performance during semantic perception task. Context information can provide rich scene category priors. To aggregate adaptive semantic contexts to pixels, we design a new image encoder

integrating Context-Guided Attention (CGA), Spatial Context Alignment (SCA) and Pixel-Point Refinement (PPR).

Context-Guided Attention. The CGA is used to enhance semantic discrimination. The core of CGA lies in employing pixel-context attention as guidance to aggregate contexts for each pixel. Given features $\mathbf{F}_{\text{c}}^{32} \in \mathbb{R}^{8C\times\frac{H}{32}\times\frac{W}{32}}$, we first employs a convolutional layer and reshape operation to generate $\mathbf{Q} \in \mathbb{R}^{N\times 2C}$. Meanwhile, feature $\mathbf{F}_{\text{c}}^{32}$ is flattened into contexts $\mathbf{C} \in \mathbb{R}^{8C\times M}$ (where M is the number of contexts) from multiple pooling layers, and further delivered into two Squeeze-and-Excitation layers to produce initial context representations $\mathbf{K} \in \mathbb{R}^{2C\times M}$ and $\mathbf{V} \in \mathbb{R}^{8C\times M}$, respectively. Unlike computing pixel-pixel similarity, we utilize new pixel-context attention to aggregate more semantic-dependent contexts for each pixel. These context representations are embedded with matrix multiplication and a softmax layer to generate pixel-context attention $\mathcal{A}$ by:

$$\mathcal{A} = \frac{\exp(\mathbf{Q}_i \cdot \mathbf{K}_j)}{\sum_{j=1}^{M}\exp(\mathbf{Q}_i \cdot \mathbf{K}_j)} \in \mathbb{R}^{N\times M} \tag{1}$$

Then, we aggregate semantics $\mathbf{M}_{\text{A}} \in \mathbb{R}^{8C\times\frac{H}{32}\times\frac{W}{32}}$ through matrix multiplication on $\mathcal{A}$ and $\mathbf{V}$ and reshape operation. In the end, a residual structure $\mathcal{F}_{\text{res}}$ with tanh activation and convolution layer is exploited to obtain fine-grained semantic contexts $\mathbf{F}_{\text{s}}^{32}$, which suppresses redundant features and enhances salient ones (*e.g.,* boundaries):

$$\mathbf{F}_{\text{s}}^{32} = \mathbf{F}_{\text{c}}^{32} + \mathbf{M}_{\text{A}} \odot \mathcal{F}_{\text{res}}(\mathbf{M}_{\text{A}}) \tag{2}$$

where $\odot$ is the Hadamard production.

Spatial Context Alignment. The SCA is responsible for enhancing semantic features with low-level details by adaptive cross-level alignment. A straightforward solution (Lin et al. 2017a,b) is to adopt simple multi-level feature fusion, which weakens the discrimination of the overall features. In contrast, SCA utilizes a feature-group mechanism into multiple sub-features for spatial alignment and fusion. Since cross-level fusion is beneficial within the same dimension, we first transforms high-dimension features $\mathbf{F}_{\text{s}}^{32}$ into unified channels with low-dimension $\mathbf{F}_{\text{c}}^{16}$ through a convolution and upsampling layer. Next, a spatial-group operation to remedy the loss of spatial details. The spatio-group takes the concatenated feature $\mathbf{F} \in \mathbb{R}^{8C\times\frac{H}{16}\times\frac{W}{16}}$ of $\mathbf{F}_{\text{s}}^{32}$ and $\mathbf{F}_{\text{c}}^{16}$ as input. These features are then split into two parts and processed through a convolution block $\mathcal{F}_{\text{cb}}$ with multiple groups g, each with dimension $\frac{4C}{g}$, for predicting offsets $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$:

$$\boldsymbol{\eta}_1[:g,:,:], \boldsymbol{\eta}_2[g:,:,:] = \mathcal{F}_{\text{cb}}(\text{Split}(\mathbf{F})) \tag{3}$$

where $\mathcal{F}_{\text{cb}}$ with 1×1 Conv, BN, ReLU and 3×3 Conv layer.

Further, to achieve cross-level feature fusion adaptively, a gate mechanism $\mathcal{F}_{\text{gate}}$ is used to bridge representation gaps. Meanwhile, the features are aligned using an alignment function $\mathcal{F}_{\text{alg}}$. The above process can be calculated as:

$$\begin{gathered}\boldsymbol{\beta}_h, \boldsymbol{\beta}_l = \mathcal{F}_{\text{gate}}(\text{Split}(\mathbf{F})) \\ \mathbf{F}_{\text{f}}^{16} = \boldsymbol{\beta}_h \odot \mathcal{F}_{\text{alg}}(\mathbf{F}_{\text{s}}^{32}, \boldsymbol{\eta}_1) + \boldsymbol{\beta}_l \odot \mathcal{F}_{\text{alg}}(\mathbf{F}_{\text{c}}^{16}, \boldsymbol{\eta}_2) \\ \mathcal{F}_{\text{alg}}(F, \eta) = \sum\nolimits_{h\in H}^{H}\sum\nolimits_{w\in W}^{W} F_{h,w}|h-\eta_h||w-\eta_w|\end{gathered} \tag{4}$$

where $\mathcal{F}_{\text{gate}}(x) = 1 + \tanh(x)$; h and w represent pixel position; $\boldsymbol{\beta_h}$, $\boldsymbol{\beta_l}$ are two gate masks. Finally, multi-level fusion $\mathbf{F}_{\text{f}}^{4}$ is generated through multiple alignment operations.

Point-Pixel Refinement. The PPR aims to establish point-to-pixel correspondence for refining image-modal features. We employ the LiDAR-to-camera projection that transforms 3D point $\mathbf{p}_i = (x_i, y_i, z_i)$ into 2D position $\mathbf{c}_i = (u_i, v_i)$ in the image plane through camera intrinsic parameters and vehicle parameters. Since only a portion of the LiDAR points are within the image view, we filter out points outside the image using binary masks, setting the mask to 1 for points within the image. This allows us to obtain each point-to-pixel mapping $\mathbf{M}_i^{p\to c}$, which is used to refine pixel features based on the projected point index.

Semantic-Geometry Transformer

A key insight behind the panoptic segmentation task is that it could capture the fine-grained details of perceived objects in the scene, such as the semantic and geometric properties. To do so, a novel Semantic-Geometry Transformer is designed to generate semantic-geometry panoptic representations, which involves: geometry aggregation, semantic learning and semantic-geometry fusion.

Geometry Aggregation. Based on the point-to-pixel mapping $\mathbf{M}^{p\to c}$, we firstly transform the image-modal $\mathbf{F}_{\text{f}}^{4}$ to obtain 2D point-wise features $\mathbf{F}_{\text{i}\to\text{p}} \in \mathbb{R}^{N_p\times C_c}$. Similarly, using voxel-to-point mapping through nearest interpolation on the voxel-model $\mathbf{F}_{\text{v}}$, we can obtain 3D point-wise features $\mathbf{F}_{\text{v}\to\text{p}} \in \mathbb{R}^{N_p\times C_v}$. Then, we fuse these as point-pixel geometry representations $\mathbf{G}_{\text{f}}$, which can be formulated as:

$$\mathbf{G}_{\text{f}} = \mathcal{F}_{\text{linear}}(\mathbf{F}_{\text{i}\to\text{p}} © \mathbf{F}_{\text{v}\to\text{p}}) \in \mathbb{R}^{N_p\times C} \tag{5}$$

where $\mathcal{F}_{\text{linear}}$ is a linear layer with C-dimension output.

Semantic Learning. The SGTransformer first builds two MLP-based semantic classifier $\mathbb{S}_{\text{2D}}$ and $\mathbb{S}_{\text{3D}}$ to learn semantic features from image-modal and voxel-modal, and perform spatial softmax to generate the probability distributions of semantic categories, which can be calculated as follows:

$$\mathbf{P}_{\text{c}} = \mathcal{F}_{\text{softmax}}(\mathbb{S}_{\text{2D}}(\mathbf{F}_{\text{f}}^{4}))\ ,\ \mathbf{P}_{\text{v}} = \mathcal{F}_{\text{softmax}}(\mathbb{S}_{\text{3D}}(\mathbf{F}_{\text{v}})) \tag{6}$$

where $\mathbf{P}_{\text{c}} \in \mathbb{R}^{(H_c\times W_c)\times N_s}$ and $\mathbf{P}_{\text{v}} \in \mathbb{R}^{N_v\times N_s}$ represent the semantic category probabilities from image-modal and voxel-modal, respectively; N_s are the number of semantic categories. To further enhance scene category priors, we divide the categories into foreground and background $\mathbf{P}_{\text{c}} = \{\mathbf{P}_{\text{fg}}, \mathbf{P}_{\text{bg}}\}$. The $\mathbf{P}_{\text{fg}}$ is filtered by a threshold $\boldsymbol{\delta}$ to refine the foreground prediction, represented by:

$$\mathbf{P}_{\text{fg}} = \mathcal{M}_{\text{fg}} \cdot \mathbf{P}_{\text{fg}},\ \ \mathcal{M}_{\text{fg}} = \begin{cases}1,\ \mathbf{P}_{\text{fg}} \geq \boldsymbol{\delta} \\ 0,\ \mathbf{P}_{\text{fg}} < \boldsymbol{\delta}\end{cases} \tag{7}$$

Then, we update the $\mathbf{P}_{\text{c}}$ based on refined foreground distribution. Similarly, the $\mathbf{P}_{\text{v}}$ is also updated. Next, these with rich category priors are embedded as semantics, denoted as $\mathbf{E}_{\text{v}} = \mathbf{P}_{\text{v}} \otimes \mathbf{F}_{\text{v}} \in \mathbb{R}^{N_s\times C_v}$ and $\mathbf{E}_{\text{c}} = \mathbf{P}_{\text{c}} \otimes \mathbf{F}_{\text{f}}^{4} \in \mathbb{R}^{N_s\times C_c}$. Meanwhile, we train the semantic classifiers using cross-entropy (CE) loss through the point-to-pixel and voxel-to-point supervision to refine category prediction:

$$\mathcal{L}_C = \text{CE}(\mathbf{Y}_{\text{3D}}, \hat{\mathbf{Y}}_{\text{2D}}) + \text{CE}(\mathbf{Y}_{\text{3D}}, \hat{\mathbf{Y}}_{\text{3D}}) \tag{8}$$

Method	Venue	PQ	PQ†	SQ	RQ	PQ^{th}	SQ^{th}	RQ^{th}	PQ^{st}	SQ^{st}	RQ^{st}	mIoU
GP-S3Net (Razani et al. 2021)	ICCV'21	61.0	67.5	84.1	72.0	56.0	85.3	65.2	66.0	82.9	78.7	75.8
EfficientLPS (Sirohi et al. 2021)	TRO'21	62.0	65.6	83.4	73.9	56.8	83.2	68.0	70.6	83.8	83.6	65.6
Panoptic-PolarNet (Zhou, Zhang, and Foroosh 2021)	CVPR'21	67.7	71.0	86.0	78.1	65.2	87.2	74.0	71.9	83.9	84.9	69.3
SCAN (Xu et al. 2022)	AAAI'22	65.1	68.9	85.7	75.3	60.6	85.7	70.2	72.5	85.7	83.8	77.4
Panoptic PH-Net (Li et al. 2022)	CVPR'22	74.7	77.7	88.2	84.2	74.0	89.0	82.5	75.9	86.8	86.9	79.7
CPSeg HR (Li et al. 2023a)	ICRA'23	71.1	75.6	85.5	82.5	71.5	87.3	81.3	70.6	83.6	83.7	73.2
4D-Former (Athar et al. 2023)	CoRL'23	77.3	80.9	89.0	86.5	79.6	90.4	87.8	73.5	86.7	84.1	78.9
PUPS (Su et al. 2023)	AAAI'23	74.7	77.3	89.4	83.3	75.4	91.8	81.9	73.6	85.3	85.6	-
LCPS (Zhang et al. 2023)	ICCV'23	79.8	84.0	89.8	88.5	82.3	91.7	89.6	75.6	86.7	86.5	80.5
CFNet (Li et al. 2023b)	CVPR'23	75.1	78.0	88.8	84.6	74.8	89.8	82.9	76.6	87.1	**87.3**	79.3
SGFormer (Ours)	AAAI'25	**80.9**	**85.1**	**90.4**	**89.2**	**83.3**	**92.2**	**90.1**	**76.7**	**87.3**	87.2	**81.3**

Table 1: Comparison of 3D panoptic segmentation on nuScenes validation set, in which PQ% is the primary metric for comparison. The first- and second-best results are highlighted in **bold** and underline, respectively.

Method	PQ	PQ†	SQ	RQ	mIoU
EfficientLPS (Sirohi et al. 2021)	62.4	66.0	83.7	74.1	66.7
SPVNAS + CenterPoint (Tang et al. 2020)	72.2	76.0	88.5	81.2	76.9
AF2S3Net + CenterPoint (Cheng et al. 2021)	76.8	80.6	89.5	85.4	78.8
Panoptic PH-Net (Li et al. 2022)	80.1	82.8	**91.1**	87.6	80.2
CPSeg (Li et al. 2023a)	73.2	76.3	82.7	88.1	73.7
4D-Former (Athar et al. 2023)	78.0	-	89.7	86.6	80.4
LCPS (Zhang et al. 2023)	79.5	82.3	90.3	87.7	78.9
SGFormer (Ours)	**80.4**	**83.1**	90.8	**88.2**	**80.7**

Table 2: Comparison on nuScenes test set.

Method	PQ	PQ†	SQ	RQ	mIoU
DS-Net (Hong et al. 2021)	57.7	63.4	77.6	68.0	63.5
EfficientLPS (Sirohi et al. 2021)	59.2	65.1	75.0	69.8	64.9
Panoptic PH-Net (Li et al. 2022)	61.7	-	-	-	65.7
4D-Former (Athar et al. 2023)	60.7	65.4	76.0	70.3	66.3
PUPS (Su et al. 2023)	**64.4**	68.6	81.5	**74.1**	-
CFNet (Li et al. 2023b)	62.7	67.5	-	-	67.4
LCPS (Zhang et al. 2023)	59.0	68.8	79.8	68.9	63.2
SGFormer (Ours)	64.2	**73.3**	**81.8**	73.6	**68.1**

Table 3: Comparison on SemanticKITTI validation set.

where $\hat{\mathbf{Y}}_{2D}$ and $\hat{\mathbf{Y}}_{3D}$ denote the predicted pixel and voxel label, respectively; $\mathbf{Y}_{3D}$ is the ground-truth point label.

Semantic-Geometry Fusion. The semantic-category embeddings $\mathbf{E}_v$ and $\mathbf{E}_c$ possess rich semantic contexts, while the point-pixel geometry features $\mathbf{G}_f$ contain geometry-consistency information. To integrate these features effectively for fine-grained panoptic representations, a semantic-geometry attention block is introduced, utilizing the attention mechanism to model potential semantic interactions through Multi-Head Self-Attention (MHSA) and facilitate semantic-geometry fusion via Multi-Head Cross-Attention (MHCA). To be specific, the self-attention layer takes the concatenated semantic feature $\mathbf{F}_e \in \mathbb{R}^{2N_s \times C}$, formed by projecting $\mathbf{E}_v$ and $\mathbf{E}_c$ as inputs. For each head in the multi-head self-attention layer, three learnable matrices $\mathbf{W}^Q$, $\mathbf{W}^K$ and $\mathbf{W}^V$ project the $\mathbf{F}_e$ to query $\mathbf{Q}_s \in \mathbb{R}^{2N_s \times d}$, key $\mathbf{K}_s \in \mathbb{R}^{2N_s \times d}$ and value $\mathbf{V}_s \in \mathbb{R}^{2N_s \times d}$, where $d = C/h$ and h is the number of heads. Subsequently, the attention matrix $\mathbf{A}_s$ is calculated by applying a row-wise softmax function on $\mathbf{Q}_s\mathbf{K}_s^T$. The multi-modal semantic relationships $\mathbf{S}_f$ are formulated as $\mathbf{A}_s\mathbf{V}_s$, capturing complex semantic interactions between the image and LiDAR modalities.

$$\mathbf{S}_f = \mathcal{F}_{\text{Norm}}(\mathcal{F}_{\text{MHSA}}(\mathbf{F}_e) + \mathbf{F}_e) \in \mathbb{R}^{2N_s \times C} \tag{9}$$

Method	PQ	PQ†	SQ	RQ	mIoU
TORNADO-Net (Gerdzhev et al. 2021)	33.7	43.3	68.4	46.0	44.5
DS-Net (Hong et al. 2021)	35.6	45.9	68.6	49.2	54.5
GP-S3Net (Razani et al. 2021)	48.7	60.3	61.3	63.7	**61.8**
CFNet† (Li et al. 2023b)	50.4	61.6	70.2	66.8	60.9
LCPS† (LiDAR) (Zhang et al. 2023)	47.1	58.7	66.2	64.1	57.4
SGFormer (LiDAR)	**51.7**	**64.2**	**71.4**	**68.2**	61.5

Table 4: Comparison on SemanticPOSS validation set.

Method	Stuck (20%)		Occlusion		FoV (-$2\pi/3, 2\pi/3$)	
	PQ	mIoU	PQ	mIoU	PQ	mIoU
Panoptic-PolarNet†	53.6	55.9	-	-	41.1	49.2
CFNet†	60.7	65.0	-	-	49.4	56.3
LCPS†	66.9	68.5	76.5	77.9	55.7	61.3
SGFormer	**69.2**	**70.3**	**78.9**	**79.9**	**59.1**	**63.9**
Improvement	↑2.3	↑1.8	↑2.4	↑2.0	↑3.4	↑2.6

Table 5: Competitive results on different robustness setting. † means the performance is reported by its official code.

Then, we perform semantic-geometry fusion on the $\mathbf{S}_f$ and $\mathbf{G}_f$ using MHCA to generate the semantic-geometry panoptic representations. The cross-attention layer takes the geometry features $\mathbf{G}_f$ and semantic relationships $\mathbf{S}_f$ as inputs. Unlike self-attention, the query $\mathbf{Q}_g$ is derived from the projection of $\mathbf{G}_f$, while the key $\mathbf{K}_r$ and value $\mathbf{V}_r$ are updated from $\mathbf{S}_f$. Thus, the attention matrix formulated as $\mathbf{Q}_g\mathbf{K}_r^T$ is to aggregate more semantics related geometric points. Finally, we employ a fully connected feed-forward network (FFN) to produce semantic-geometry fusion features, denoted as $\mathbf{F}_{sg} \in \mathbb{R}^{N_p \times C}$. The above process is expressed as:

$$\begin{aligned} \mathbf{M}_{sg} &= \mathcal{F}_{\text{Norm}}(\mathbf{G}_f + \mathcal{F}_{\text{MHCA}}(\mathbf{G}_f, \mathbf{S}_f)) \\ \mathbf{F}_{sg} &= \mathcal{F}_{\text{Norm}}(\mathbf{M}_{sg} + \mathcal{F}_{\text{FFN}}(\mathbf{M}_{sg})) \end{aligned} \tag{10}$$

Panoptic Network

Following (Zhou, Zhang, and Foroosh 2021), we input multi-modal $\mathbf{F}_{sg}$ into BEV features $\mathbf{F}_{sg}^{BEV}$ through the grid index of points and BEV pooling. Then, the panoptic network is divided into semantic and instance heads.

Semantic Head. Based on the predicted logits of each point on semantic labels $\hat{\mathbf{S}}$ and ground-truth $\mathbf{S}$, we could get semantic loss $\mathcal{L}_S$ by cross-entropy loss.

Instance Head. Based on the predicted center offset $\hat{\mathbf{o}} \in$

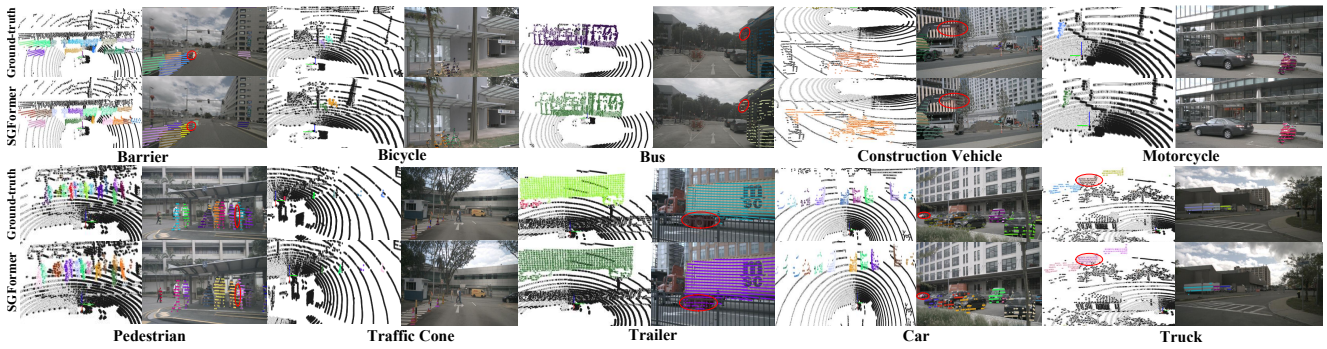


Figure 4: Class-wise qualitative visualization on nuScenes validation set. Best viewed with zoom and color. Red circles demonstrate that our SGFormer performs better in many details than ground-truth labels.

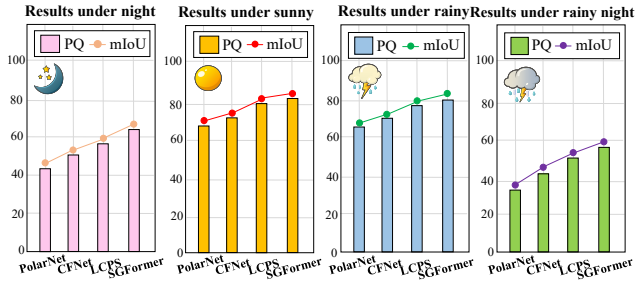


Figure 5: Comparison under different scene conditions.

$\mathbb{R}^{H\times W\times 2}$ and ground-truth $\mathbf{o}$, the L1 loss ($\mathcal{L}_o$) for optimizing offset regression. Meanwhile, given the predicted center heatmap $\hat{\mathbf{H}} \in \mathbb{R}^{H\times W\times 1}$ and ground-truth $\mathbf{H}$, we adopt MSE loss ($\mathcal{L}_{hm}$) to optimize heatmap regression. The total loss of instances $\mathcal{L}_I$ is formulated as:

$$\mathcal{L}_I = \lambda_{hm}\mathcal{L}_{hm} + \lambda_o\mathcal{L}_o \tag{11}$$

Fianlly, a hybrid loss is used to supervise the training process:

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_I + \lambda_c\mathcal{L}_C \tag{12}$$

Experiments

Experiment Setup

Datasets. nuScenes (Fong et al. 2022) is a large-scale benchmark, containing 1000 scenes. SemanticKITTI (Behley et al. 2019) is an outdoor dataset, consisting of 22 sequences. SemanticPOSS (Pan et al. 2020) is a challenging benchmark, including 2988 scenes of 6 sequences.

Metrics. We evaluate performance using panoptic, segmentation and recognition quality (PQ, SQ, RQ), with separate metrics for stuff (*e.g.,* PQ^{st}) and things (*e.g.,* PQ^{th}) classes. Mean IoU (mIoU) is used for semantic segmentation. The $\text{PQ}^{\dagger}$ denotes PQ of stuff classes is replaced by their IoU.

Implementation Details. The specific details are provided in the supplementary material. In ASCA, the groups $g = 8$ for alignment. In SGTransformer, we set $\boldsymbol{\delta}$ to 0.1 and use two fusion layers, each layer with four self-attention and one cross-attention equipped with 128 input channels. In terms of loss weights, we set $\lambda_{hm} = 100$, $\lambda_o = 10$ and $\lambda_c = 1$.

ResNet	ASCA	Naive fusion	SGTransformer	2×	3×	PQ	mIoU
✔		✔				74.6	76.5
✔	✔	✔				77.1	78.2
✔			✔	✔		78.6	79.4
✔	✔		✔	✔		**80.9**	**81.3**
✔	✔		✔		✔	80.6	81.0

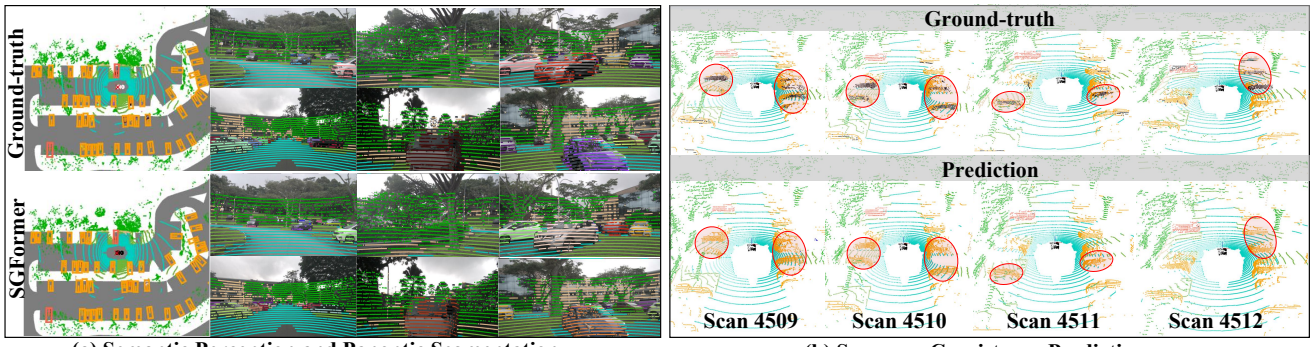
Table 6: Ablation study of network architecture. **2**× and **3**×: the number of semantic-geometry fusion layer.

Comparison with State-of-the-Arts

Results on nuScenes. As shown in Table 1, SGFormer outperforms state-of-the-arts with higher panoptic segmentation performance on the nuScenes val set. Specifically, our method surpasses recent LCPS (Zhang et al. 2023) by 1.1% on PQ and 0.8% on mIoU. Moreover, in Table 2, our SGFormer achieves top-performing results than Panoptic-PHNet (Li et al. 2022) and further surpasses LCPS on all metrics. These results demonstrate SGFormer can better distinguish objects through semantic-geometry fusion, significantly advancing 3D panotic segmentation.

Results on SemanticKITTI and SemanticPOSS. Note that SemanticKITTI has only two cameras in the front view, resulting in fewer points matched with images. Results in Table 3 reveal that SGFormer not only performs competitively with state-of-the-art approach PUPS (Su et al. 2023) but also outperforms recent LiDAR-camera fusion method with an obvious 5.2% PQ improvement. Additionally, on SemanticPOSS, which features much smaller and sparser point clouds, SGFormer surpasses existing methods across almost all metrics in Table 4. These results highlight our SGFormer in leveraging semantic information and spatial geometry.

Robustness towards Noisy data and Outliers. We adopt similar ways in nuScenes-R (Yu et al. 2023) to introduce noisy data and outliers in the point clouds and images, including camera occlusion, LiDAR frames 20% stuck (Stuck) and point with limited Field-of-View (FoV). As depicted in Table 5, SGFormer outperforms leading methods, showing a 3.4% PQ improvement on LiDAR FOV and a 2.4% PQ gain on camera occlusion compared to LCPS. These results



(a) Semantic Perception and Panoptic Segmentation (b) Sequence Consistency Prediction

Figure 6: Qualitative visualization on nuScenes. (a) illustrates semantic perception and panoptic sgemantion among the ground-truth and SGFormer. (b) represents sequence consistency segmentation among the ground-truth and SGFormer.

Method	PQ	mIoU
w/o Context-Guided Attention (CGA)	80.0	80.6
w/o Spatial Context Alignment (SCA)	79.8	80.5
w/o Point-Pixel Refinement (PPR)	80.6	81.1
Spatial-Channel Enhancement	80.3	80.8
Context-Guided Attention	**80.9**	**81.3**
3DCVF (Yoo et al. 2020)	80.3	80.7
LiraFusion (Song, Zhao, and Skinner 2024)	80.6	81.0
Spatial Context Alignment	**80.9**	**81.3**

Table 7: Detailed ablation study for the ASCA.

Method	PQ	mIoU
w/o Geometry Aggregation	80.0	80.6
w/o Semantic Learning	79.6	80.2
Feature Concatenation	77.1	78.2
Cross-Attention	78.2	79.1
Semantic-Geometry Fusion	**80.9**	**81.3**

Table 8: Ablation study for the SGTransformer.

demonstrate the superior robustness of our method.

Adaptation to Different Scenes. Following the scene description in nuScenes, we categorize conditions into four scenes: sunny daytime, rainy day, night, and rainy night. As detailed in Fig. 5, SGFormer outperforms existing methods with higher results under changing weather scenarios. Particularly, during nights and rainy nights, the visibility of both the LiDAR and camera is compromised, our model performs more precise predictions, demonstrating that the adaptability of proposed method to different scene conditions.

Ablation Studies

Analysis of Network Architecture. As shown in Table 6, introducing ASCA in Row-2 yields a 2.5% PQ and 1.7% mIoU improvement, indicating the significance of adaptive semantic context aggregation. Meanwhile, replacing naive fusion with SGTransformer in Row-3 results in a 4.0% PQ and 2.9% mIoU boost. Moreover, combining the proposed modules with two semantic-geometry fusion layers further enhances performance. However, in the last row, increasing the number of fusion layers leads to a 0.3% decline in PQ due to redundant iterations discarding some semantic and geometric features necessary for 3D panoptic segmentation.

Effectiveness of ASCA. As shown in Table 7, experiments in the Row-1, 2 and 3 verify each component contributes to performance improvement. Moreover, we replace CGA with spatial-channel enhancement and also modify the input of gated fusion from 3DCVF (Yoo et al. 2020) and LiraFusion (Song, Zhao, and Skinner 2024) as replacement for SCA to evaluate their abilities. As detailed in Row-4, 6 and 7, it is evident that our CGA and SCA deliver better results.

Effectiveness of SGTransformer. As shown in Table 8, the results in Rows 1 and 2 demonstrate the necessity of geometry aggregation and semantic learning through aggregating spatial geometry and embedding semantic priors. Further, comparing the results of Row-3, 4 and 5, SGFusion can achieve better performance than using a simple feature concatenation and cross-attention layer to fuse multi-modal features, which show the effectiveness of our fusion method.

Further Discussion

Generalization Ability. In Table 9, ASCA is employed as a plug-and-play module to enhance baselines DDRNet (Pan et al. 2022) and PIDNet (Xu, Xiong, and Bhattacharyya 2023) on semantic tasks like Cityscapes (Cordts et al. 2016), PASCAL (Everingham et al. 2010) and CamVid (Brostow, Fauqueur, and Cipolla 2009). We further assess generalization on out-of-distribution data using two corruption datasets from the Robo3D (Kong et al. 2023) benchmark. As detailed in Table 10, SGFormer achieves superior results, demonstrate the strong generalization ability of our method.

Analysis of Semantic and Geometry. In Fig. 7, the left represents using different semantic labels for class-wise segmentation. It is evident that introducing semantic information (w *vs.* w/o) enhances performance. While replacing the ground-truth in semantic learning with our semantic prediction achieves close performance, demonstrating our model can generate accurate semantic information. Meanwhile, the bar graph on the right (points at varying distance ratio from the instance center) for class-wise classification shows the

Method	Cityscapes(*val*)	PASCAL VOC(*val*)	CamVid(*test*)
DDRNet-23†	79.3	58.2	80.4
DDRNet-23*	**80.0** ↑ 0.7	**59.2** ↑ 1.0	**81.0** ↑ 0.6
PIDNet-M†	79.9	57.9	80.0
PIDNet-M*	**80.5** ↑ 0.6	**58.7** ↑ 0.8	**80.5** ↑ 0.5

Table 9: Comparison (mIoU) on different baselines with using ASCA (*). † denotes our re-implementation.

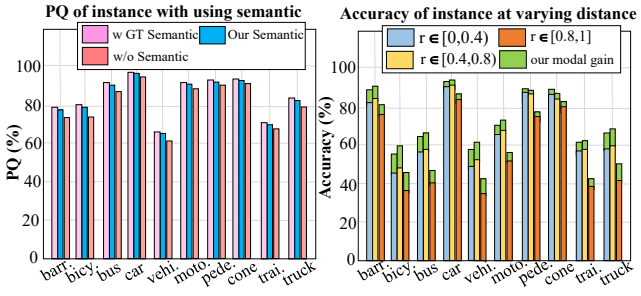


Figure 7: Left: The PQ per-class of instance with using different semantic labels. Right: The class-wise classification accuracy of instance at varying distance.

greatest accuracy increase, suggesting that our model can effectively capture spatial geometry to improve perception.

Model Efficiency and Accuracy. In Fig. 8, we report the results of PQ to represent the accuracy of different methods and latency (ms) to represent the efficiency of the models. Our SGFormer can maintain lower complexity and latency than LCPS, while achieving higher accuracy. Additionally, SGFormer has real-time runtime in 100 ms boundary and similar latency with LiDAR-based Panoptic PH-Net.

Qualitative Analysis

As shown in Fig. 4, SGFormer achieves accurate class-wise segmentation on small and rare objects, even outperforming ground-truth. SGFormer also conducts effective semantic perception and panoptic segmentation in Fig. 6 (a), while delivering robust prediction in a sequence of frames from moving ego-car (Fig. 6 (b)) with consistent segmentation. These results demonstrate our semantic-geometry fusion can generate fine-grained panoptic predictions. More details refer to the supplementary material for analyses.

Conclusion

In this paper, we exploit semantic contexts and spatial geometry for 3D panoptic segmentation. To this end, we propose SGFormer, a semantic-geometry fusion transformer to generate semantic-geometry panoptic representations, which includes: an adaptive semantic-context aggregation and a semantic-geometry transformer. Comparisons and ablations demonstrate the effectiveness of our method.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant (U24A20242, U2033210, 61906168 and 62272267) and in part by the Zhejiang Provincial Natural Science Foundation under Grant (LY23F020023, LDT23F02024F02 and LZ23F020001), the Hangzhou AI major scientific and technological innovation project under Grant (2022AIZD0061) and Anhui Key Laboratory of Bionic Sensing and Advanced Robot Technology Proiect (AHFS2024KF04).

Method	SemanticKITTI-C mCE ↓	SemanticKITTI-C mRR ↑	nuScenes-C mCE ↓	nuScenes-C mRR ↑
Cylinder3D (Zhu et al. 2021)	103.3	80.1	111.8	72.9
2DPASS (Yan et al. 2022)	106.1	77.5	98.6	75.2
WaffleIron (Puy, Boulch, and Marlet 2023)	109.5	72.2	106.7	72.8
SGFormer	**96.8**	**81.4**	**93.7**	**82.1**

Table 10: Comparison on out-of-distribution generalization.

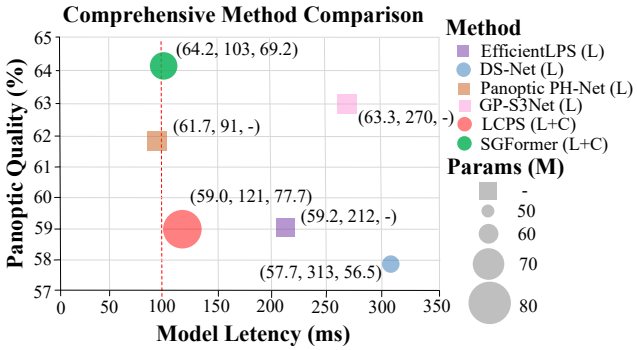


Figure 8: Efficiency *vs.* accuracy on SemanticKITTI.

References

Athar, A.; Li, E.; Casas, S.; and Urtasun, R. 2023. 4D-Former: Multimodal 4D Panoptic Segmentation. In *Conference on Robot Learning*, 2151–2164.

Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; and Tai, C.-L. 2022. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1090–1099.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9297–9307.

Brostow, G. J.; Fauqueur, J.; and Cipolla, R. 2009. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2): 88–97.

Cheng, R.; Razani, R.; Taghavi, E.; Li, E.; and Liu, B. 2021. 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*, 12547–12556.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3213–3223.

Dong, G.; Yan, Y.; Shen, C.; and Wang, H. 2020. Real-time high-performance semantic image segmentation of urban street scenes. *IEEE Transactions on Intelligent Transportation Systems*, 22(6): 3258–3274.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88: 303–338.

Fong, W. K.; Mohan, R.; Hurtado, J. V.; Zhou, L.; Caesar, H.; Beijbom, O.; and Valada, A. 2022. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7(2): 3795–3802.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3146–3154.

Gerdzhev, M.; Razani, R.; Taghavi, E.; and Bingbing, L. 2021. Tornado-net: multiview total variation semantic segmentation with diamond inception module. In *2021 IEEE International Conference on Robotics and Automation*, 9543–9549.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Hong, F.; Zhou, H.; Zhu, X.; Li, H.; and Liu, Z. 2021. Lidar-based panoptic segmentation via dynamic shifting network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13090–13099.

Huang, Z.; Wei, Y.; Wang, X.; Liu, W.; Huang, T. S.; and Shi, H. 2021. Alignseg: Feature-aligned segmentation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 550–557.

Kirillov, A.; He, K.; Girshick, R.; Rother, C.; and Dollár, P. 2019. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9404–9413.

Kong, L.; Liu, Y.; Li, X.; Chen, R.; Zhang, W.; Ren, J.; Pan, L.; Chen, K.; and Liu, Z. 2023. Robo3d: Towards robust and reliable 3d perception against corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19994–20006.

Li, E.; Razani, R.; Xu, Y.; and Liu, B. 2023a. Cpseg: Cluster-free panoptic segmentation of 3d lidar point clouds. In *2023 IEEE International Conference on Robotics and Automation*, 8239–8245.

Li, J.; He, X.; Wen, Y.; Gao, Y.; Cheng, X.; and Zhang, D. 2022. Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering pseudo heatmap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11809–11818.

Li, X.; Zhang, G.; Wang, B.; Hu, Y.; and Yin, B. 2023b. Center Focusing Network for Real-Time LiDAR Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13425–13434.

Lin, G.; Milan, A.; Shen, C.; and Reid, I. 2017a. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1925–1934.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017b. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D. L.; and Han, S. 2023. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE International Conference on Robotics and Automation*, 2774–2781.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Lv, Q.; Sun, X.; Chen, C.; Dong, J.; and Zhou, H. 2021. Parallel complement network for real-time semantic segmentation of road scenes. *IEEE Transactions on Intelligent Transportation Systems*, 23(5): 4432–4444.

Pan, H.; Hong, Y.; Sun, W.; and Jia, Y. 2022. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. *IEEE Transactions on Intelligent Transportation Systems*, 24(3): 3448–3460.

Pan, J.; Wang, Z.; and Wang, L. 2024. Co-occ: Coupling explicit feature fusion with volume rendering regularization for multi-modal 3d semantic occupancy prediction. *IEEE Robotics and Automation Letters*.

Pan, Y.; Gao, B.; Mei, J.; Geng, S.; Li, C.; and Zhao, H. 2020. Semanticposs: A point cloud dataset with large quantity of dynamic instances. In *2020 IEEE Intelligent Vehicles Symposium*, 687–693.

Peng, J.; Liu, Y.; Tang, S.; Hao, Y.; Chu, L.; Chen, G.; Wu, Z.; Chen, Z.; Yu, Z.; Du, Y.; et al. 2022. Pp-liteseg: A superior real-time semantic segmentation model. *arXiv preprint arXiv:2204.02681*.

Puy, G.; Boulch, A.; and Marlet, R. 2023. Using a waffle iron for automotive point cloud semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3379–3389.

Razani, R.; Cheng, R.; Li, E.; Taghavi, E.; Ren, Y.; and Bingbing, L. 2021. Gp-s3net: Graph-based panoptic sparse semantic segmentation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16076–16085.

Sirohi, K.; Mohan, R.; Büscher, D.; Burgard, W.; and Valada, A. 2021. Efficientlps: Efficient lidar panoptic segmentation. *IEEE Transactions on Robotics*, 38(3): 1894–1914.

Song, J.; Zhao, L.; and Skinner, K. A. 2024. Lirafusion: Deep adaptive lidar-radar fusion for 3d object detection. *arXiv preprint arXiv:2402.11735*.

Su, S.; Xu, J.; Wang, H.; Miao, Z.; Zhan, X.; Hao, D.; and Li, X. 2023. Pups: Point cloud unified panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2339–2347.

Tang, H.; Liu, Z.; Zhao, S.; Lin, Y.; Lin, J.; Wang, H.; and Han, S. 2020. Searching efficient 3d architectures with sparse point-voxel convolution. In *Proceedings of the European Conference on Computer Vision*, 685–702.

Xiao, Z.; Zhang, W.; Wang, T.; Loy, C. C.; Lin, D.; and Pang, J. 2023. Position-Guided Point Cloud Panoptic Segmentation Transformer. *arXiv preprint arXiv:2303.13509*.

Xu, J.; Xiong, Z.; and Bhattacharyya, S. P. 2023. PIDNet: A real-time semantic segmentation network inspired by PID controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19529–19539.

Xu, S.; Wan, R.; Ye, M.; Zou, X.; and Cao, T. 2022. Sparse cross-scale attention network for efficient lidar panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2920–2928.

Yan, J.; Liu, Y.; Sun, J.; Jia, F.; Li, S.; Wang, T.; and Zhang, X. 2023. Cross modal transformer: Towards fast and robust 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18268–18278.

Yan, X.; Gao, J.; Zheng, C.; Zheng, C.; Zhang, R.; Cui, S.; and Li, Z. 2022. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *Proceedings of the European Conference on Computer Vision*, 677–695.

Yang, M.; Yu, K.; Zhang, C.; Li, Z.; and Yang, K. 2018. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3684–3692.

Ye, D.; Zhou, Z.; Chen, W.; Xie, Y.; Wang, Y.; Wang, P.; and Foroosh, H. 2023. Lidarmultinet: Towards a unified multi-task network for lidar perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3231–3240.

Yoo, J. H.; Kim, Y.; Kim, J.; and Choi, J. W. 2020. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *Proceedings of the European Conference on Computer Vision*, 720–736.

Yu, K.; Tao, T.; Xie, H.; Lin, Z.; Liang, T.; Wang, B.; Chen, P.; Hao, D.; Wang, Y.; and Liang, X. 2023. Benchmarking the robustness of lidar-camera fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3188–3198.

Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; and Wang, J. 2021. OCNet: Object context for semantic segmentation. *International Journal of Computer Vision*, 129(8): 2375–2398.

Zhang, J.; and Ding, Y. 2024. Occfusion: Depth estimation free multi-sensor fusion for 3d occupancy prediction. *arXiv preprint arXiv:2403.05329*.

Zhang, Y.; Zhu, Z.; and Du, D. 2023. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9433–9443.

Zhang, Z.; Zhang, Z.; Yu, Q.; Yi, R.; Xie, Y.; and Ma, L. 2023. LiDAR-Camera Panoptic Segmentation via Geometry-Consistent and Semantic-Aware Alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3662–3671.

Zhou, Z.; Zhang, Y.; and Foroosh, H. 2021. Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13194–13203.

Zhu, X.; Zhou, H.; Wang, T.; Hong, F.; Ma, Y.; Li, W.; Li, H.; and Lin, D. 2021. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9939–9948.

Zhu, Z.; Xu, M.; Bai, S.; Huang, T.; and Bai, X. 2019. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 593–602.