Machine Intelligence Research
www.mi-research.net

22(5), October 2025, 956-968
DOI: 10.1007/s11633-024-1508-2

LiDAR-camera Cooperative Semantic Segmentation

He Guan[1,2] Chunfeng Song[1,2] Zhaoxiang Zhang[1,2,3]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China

[2] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China

[3] Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong 999077, China

Abstract: LiDAR and cameras are two prominent sources for parsing the semantics of the scene. While the former provides accurate physical measurements, it lacks the colour and texture appearance that the latter excels in. Fully exploiting the rich information of multimodal data is beneficial for comprehensive perception of the environment. To cope with the dual challenges of heterogeneity and consistency faced by multimodal features, we propose a unified multimodal cooperative segmentation workflow. By establishing cross-view cooperation paths, we achieve cross-view feature interactions and missing modality completions. The pre-synchronisation mechanism preserves the alignment semantics and geometry while decoupling the processing of multimodal data augmentation. Notably, our workflow jointly performs LiDAR-based 3D semantic segmentation and image-based 2D semantic segmentation with promising results on two public benchmarks: the SemanticKITTI dataset and the Waymo Open dataset.

Keywords: Multimodal, semantic segmentation, cooperation, interaction, completion.

Citation: H. Guan, C. Song, Z. Zhang. LiDAR-camera cooperative semantic segmentation. *Machine Intelligence Research*, vol.22, no.5, pp.956–968, 2025. http://doi.org/10.1007/s11633-024-1508-2

1 Introduction

In large-scale outdoor scene understanding, semantic segmentation is regarded as a crucial component, and it has been widely used in fields such as autonomous driving, digital cities, and service robots[1–3]. Related research over the past few years has achieved significant breakthroughs in both image-based[4–9] and LiDAR-based[10–13] segmentation tracks. However, the single-mode solution is limited by the inherent limitations of the sensor and inevitably faces challenges in complex environments. Specifically, cameras provide dense physical appearance information, but they are heavily affected by lighting and object scale. The sparse and textureless point cloud provided by LiDAR cannot capture the details of objects, but it can perceive depth accurately and over a wide range.

Obviously, the input data from camera images and LiDAR point clouds are complementary, and making full use of multimodal information is more conducive to ensuring the robustness of perception. The core challenge lies in seamlessly connecting multimodal data for collaborative sensing. An intuitive approach is to directly unify the coordinate systems of each input modality within the view frustum through camera calibration, but this approach is limited by three factors: 1) The inherent heterogeneity between sparse LiDAR points and dense RGB images imposes limitations on direct fusion, resulting in sub-optimal feature optimization for each modality. 2) Unlike LiDAR sensors with 360° surround view acquisition, the field of view (FOV) between multi-camera images needs to be coherent to avoid possible information conflicts and loss where multi-view frustums overlap. 3) The uniqueness of data augmentations for each modality leads to a lack of synchronization and consistency guarantees when extending to multiple modalities.

To address the above issues, we propose a point-centric multimodal cooperative segmentation workflow called CoSEG. Specifically, at each stage of feature extraction for each mainstream modality, cooperative pathways are used to mount pixel- and voxel-view features to point-view features to drive the assimilation of heterogeneous modalities as much as possible. Multi-camera RGB images regularize visible point clouds to improve the generalization of reasoning in unknown point cloud areas. In turn, the coherent 360° surrounding LiDAR point cloud also provides guidance for resolving overlapping conflicts and missing perspectives from multiple cameras. This mutual support relationship enables a cross-view cooperative interaction mechanism for voxel-point feature transformation within the same modality and pixel-point feature transformation between different modalities. The cross-view cooperative completion mechanism addresses issues such as missing modalities in a robust manner.

Research Article
Manuscript received on July 17, 2023; accepted on April 3, 2024; published online on February 21, 2025
Recommended by Associate Editor Jingyi Yu
Colored figures are available in the online version at https://link.springer.com/journal/11633
© Institute of Automation, Chinese Academy of Sciences and Springer-Verlag GmbH Germany, part of Springer Nature 2025

Maintaining modal synchronization and alignment during data augmentation can be challenging due to differences in multimodal data acquisition devices and the consistency of the semantic space. Thus, we implemented a multimodal pre-synchronization mechanism that allows easy decoupling into two families of image-level and point cloud-level data augmentation in an asymmetric manner. This approach enables us to flexibly embed a variety of excellent augmented operators. Importantly, this strategy does not conflict with the workflow. Therefore, any 2D or 3D semantic segmentation network can be adapted to our pipeline.

In general, the main contributions can be summarized as follows:

1) By utilizing the interaction and completion mechanism of multimodal cross-view cooperation, our proposed multimodal cooperative segmentation workflow can simultaneously perform semantic segmentation on both the LiDAR-based branch and the image-based branch, thus realizing one-stop output of multi-dimensional semantics for the same scene.

2) With the multimodal pre-synchronization mechanism, our method also provides an entry point for diverse data augmentation operators.

3) Our approach achieves state-of-the-art 3D segmentation performance on both the SemanticKITTI and Waymo benchmarks, demonstrating the effectiveness of the proposed multimodal cooperative segmentation workflow.

2 Related work

2.1 Camera-only 2D semantic segmentation

Camera-based semantic segmentation requires predicting the semantic category of each pixel in an image. Fully convolutional networks (FCN) have been widely adopted as pioneers in solving the image-based semantic segmentation task with deep network architecture. Since then, FCN architectural variants have blossomed, such as capturing global information by extending the receptive field[4, 5], multi-scale feature aggregation[14, 15], contextual information refinement[16, 17], precise localization with boundary cues[18, 19], designing attention mechanisms to emphasize or suppress features[6, 8], and even exploring advanced subnets or hyperparameters through network architecture search[7, 20]. Recently, transformer-based architectures have dramatically improved segmentation model accuracy and speed[21], while taking into account lightweight modelling requirements[9].

2.2 LiDAR-only 3D semantic segmentation

The LiDAR-only semantic segmentation track has emerged with large-scale 3D semantic segmentation benchmarks (e.g., SemanticKITTI and Waymo). This spurred the exploration of efficient representations of point cloud data, with successive proposals such as point-based multilayer perceptrons (MLPs)[22–24], convolution-wise operators[25] or pseudo-grids[26, 27]. The above schemes are limited by time-consuming sampling and grouping in terms of efficiency, and only adapt to small point sets in terms of scale, making it difficult to generalize to large-scale sparse LiDAR scenarios. Despite unavoidable information loss, alternative spatial projections for seeking 3D point clouds remain valid, including planar[28, 29], spherical[30, 31], and their combinations[32], etc. Voxel-view balances the effectiveness and efficiency of 3D data representation by storing and processing only non-empty voxels[33]. Subsequent extensions all utilize more powerful grid variants[34–36].

2.3 Multimodal 2D-3D semantic segmentation

Unimodal schemes for LiDAR point clouds can only be presented in various representations, such as voxel-point-range generalized fusion[37] and point-voxel adaptive optimizations[11]. However, further extensions to camera modalities that provide rich appearance and texture are more compelling. Intuitive multimodal fusion focuses on the unification of coordinate systems, e.g., projecting a point cloud into a perspective view to optimize the residual fusion module[38], multi-phase fusion of range and re-projected RGB images[39], integrating pixels within the ranging image backbone via calibration matrices[40], and projecting the image into point space by segmentation logic[41]. Recent explorations have utilized auxiliary modal priors and knowledge distillation to extract adaptive multimodal features[42]. In contrast, our approach verifies that the semantics of points and images can assist each other simultaneously, rather than only 2D assisting 3D in one direction.

2.4 Augmentation with multimodal

It is well known that data augmentation is crucial in perception training, but it is rarely adopted in multimodal segmentation. Directly applying modality-specific operations to multimodal sample pairs separately would destroy the correspondence consistency from scratch, so it is common to enable only a few multimodal general operators, even abandoning the enhancement of image input. Some rare precedents[43–45] related to detection also resort to synchronization to deal with the inconsistency issue in cross-modal augmentation. In fact, as long as the correspondence of multimodal sample pair is obtained in advance, most image- and point-based operators (except those with out-of-order properties) can be inserted, and this can be easily achieved through the camera calibra-

tion matrix.

3 Methodology

3.1 Preliminary

For multimodal 3D semantic segmentation, paired samples $\{\boldsymbol{P}, \boldsymbol{X}\}$ are usually captured as input. $\boldsymbol{P} \in \mathbf{R}^{N_P \times C_P}$ denotes LiDAR point cloud, where N_P and C_P are the number and attribute (e.g., 3D coordinates and reflectance) dimensions of the input points, respectively. $\boldsymbol{X} \in \mathbf{R}^{N_c \times H \times W \times 3}$ denotes the multi-camera RGB images of the N_c cameras, where H and W are the height and width of the input image, respectively.

To address the modal heterogeneity between point clouds and images, we customize parallel backbones to independently extract intra-modal features. To accommodate the unstructured nature of point-view, point features $\mathbf{F}^P \in \mathbf{R}^{N_p \times C_p}$ are extracted from a series of multi-layer perceptrons (MLPs), where C_p indicates the number of channels of the point features. Relying on the superiority of the voxel-view, $\boldsymbol{P}$ is grouped into non-empty voxels by sparse quantization and constitutes the sparse tensor input to the generic 3D backbone to extract voxel features $\mathbf{F}^V \in \mathbf{R}^{N_v \times C_v}$. N_v is the number of non-empty voxels and C_v is the number of channels of the voxel features. In addition, we launch a common 2D backbone[8, 46] to non-linearly project $\boldsymbol{X}[c_i]$ as $\mathbf{F}^X[c_i] \in \mathbf{R}^{H_x \times W_x \times C_x}$ within the c_i-th local camera, where H_x, W_x and C_x represent the height, width and the number of channels of the camera image features, respectively.

The overview pipeline of the proposed multimodal cooperative semantic segmentation is shown in Fig. 1. We jointly optimize each backbone throughout the workflow to facilitate cooperative complementation of cross-view features for more robust and accurate perception. In addition, we perform sample de-synchronization in the augmentation stage and dynamically complement features via a cross-modal feature interaction module in the training stage, thereby alleviating the reliance on strictly paired samples in multimodal fusion. In principle, 2D and 3D backbones can be flexibly replaced with mature or lightweight counterparts.

3.2 Cross-view cooperative pathway

A major bottleneck in multimodal segmentation models is the heterogeneous information barrier, i.e., the claim to combine rich color/texture details with precise physical metrics. To address this concern, we open up a point-centric multimodal cross-view cooperation pathway that binds pixel-view and voxel-view features uniformly to the point-view. With this joint optimization of cross-view cooperation, both 2D and 3D segmentation can be effectively compensated.

Calibration projection. Intuitively, the pivot of the correspondence between LiDAR points and RGB image pixels is the camera calibration matrix. For each point coordinate (x_i, y_i, z_i), the corresponding calibration pixel (c_i, u_i, v_i) projected on the c_i-th local camera can be obtained as follows:

$$[c_i, u_i, v_i, 1]^{\mathrm{T}} = \frac{K_{c_i} T_{c_i}}{z_i} [x_i, y_i, z_i, 1]^{\mathrm{T}} \tag{1}$$

where the camera intrinsic and extrinsic and matrix are denoted as $K \in \mathbf{R}^{3\times4}$ and $T \in \mathbf{R}^{4\times4}$, respectively. However, relying on the camera calibration matrix alone to obtain the correspondence between 2D and 3D is not stable enough, and a slight calibration error can result in several or even tens of pixels of deviation. It is also impractical and unnecessary to utilize additional depth assistance for 2D to 3D lifting in the segmentation task.

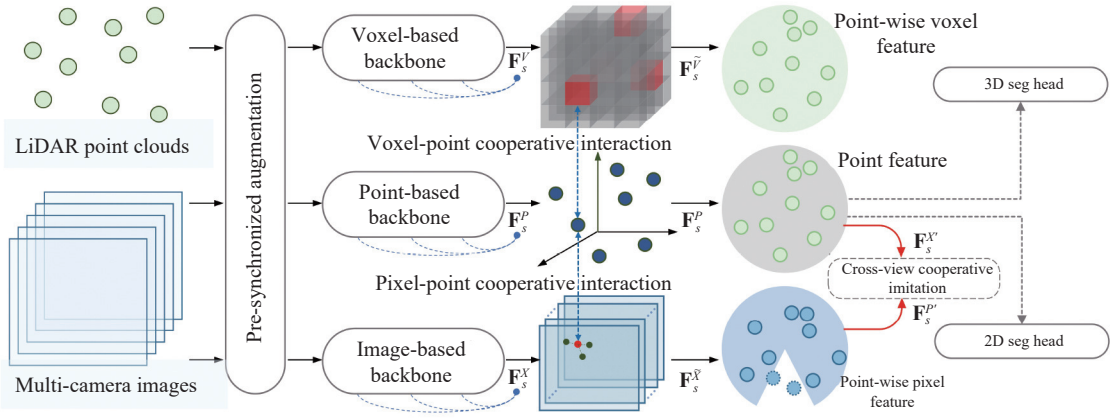


Fig. 1 Overview of the proposed multimodal cooperative segmentation workflow. We take LiDAR point clouds and multi-camera images as inputs and generate modality-specific voxel-, point- and pixel-view features through a 3D voxel encoder, a series of MLPs and a 2D image encoder respectively. Complementary information flow is achieved through cross-view cooperative interaction and completion. Additional multimodal pre-synchronization mechanisms provide an interface for various augmentations. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

Hetero-modal view cooperation for pixel-point. We mimic the top-down paradigm[47] for the pixel-point view cooperative pathway, as shown in Fig. 2. The rough correspondences between 3D reference points and 2D pixels are initialized through a pre-computed calibration and the offsets of a small group of sample points around the reference point are learned. The subsequent multi-head cross-attention module considers the local-view feature unit as the query and samples the key and the value from the supplementary view. Note that when processing the 3D side, query includes all masked 3D points; but for the 2D side, query only considers the valid sparse pixels covered by the calibrated 2D projection after rounding. Thus the transformed local-view features with cross-view cooperation are as follows:

$$\mathbf{F}^{local} = \sum_{q=1}^{Q} \mathcal{M} \times W_q \left[\sum_{k=1}^{K} A_{iqk} \times W'_q \mathbf{F}^{cross}(p_i + \Delta p_{iqk}) \right] \tag{2}$$

where $\mathbf{F}^{local}$ and $\mathbf{F}^{cross}$ represent local and cross-view features, which serve to compensate for the local point-view from the cross image-view and vice versa. The learnable weights are denoted by W_q and W'_q, while the indices q and k represent the attention head and sampling key, respectively. The attention weight A_{iqk} and sampling offset Δp_{iqk} are individually linked to the k-th sampling point in the q-th attention head. Even if the deformable mechanism is enabled, the attention computation between pixel points still has a huge overhead. To address this issue, we compress the internal channel number of the MHCA by a factor of z (where $z < 1$), and then restore the original feature shape through a feed-forward network. In addition, attaching an all-zero vector to points located in the multi-camera FOV dead zone will produce sub-optimal performance, so a binary mask $\mathcal{M}$ with N_p elements is filtered and marked as visible (marked with 1) or invisible (marked with 0).

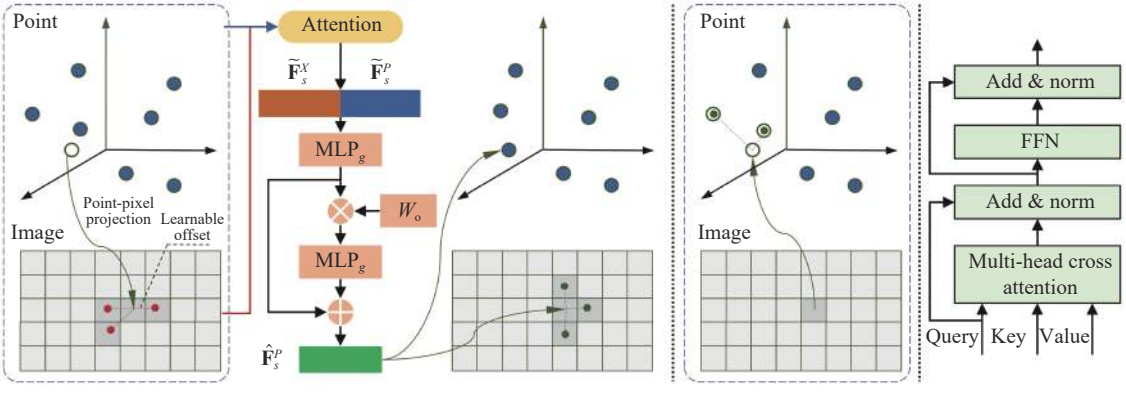
Homo-modal view cooperation for voxel-point. In contrast, the voxel-point bidirectional mapping is not reliant on the camera calibration matrix. However, it is crucial to acknowledge the presence of shifts in coordinates and mismatches in quantity between the voxel centers and the raw points. Therefore, for the voxel-to-point view pathway, we apply trilinear devoxelization to generate point-wise interpolated voxel features from the three nearest neighboring voxels. For the reverse point-to-voxel view pathway, we perform voxelization with sparse hash queries. The implementation of this part of the module relies on the torchsparse[48] wrapper function that supports efficient sparse computation.

3.3 Cross-view cooperative aggregation

Equipping such cooperation pathways at multiple scales throughout the entire workflow promotes efficient and dynamic aggregation of multimodal features. Explicitly unifying multimodal features into the same view implies that features from diverse modalities are forcibly placed in the same distribution space. This facilitates further exploration of cross-view feature interactions and view-deficient completions.

Cross-view cooperative interaction. We sequentially place the voxel-point-based homo-modal cross-view pathway and the pixel-point-based hetero-modal cross-view pathway between the backbones. The optional attention module is only adapted to the latter for learning deformable offsets (2), and its structure is depicted in Fig. 2(c).

Specifically, after the transformation of feature extraction stage s, we obtain point-wise voxel features $\mathbf{F}_s^{\widetilde{V}} \in \mathbf{R}^{N_p \times C_v}$, point-wise pixel features $\mathbf{F}_s^{\widetilde{X}} \in \mathbf{R}^{N_p \times C_x}$, and point features $\mathbf{F}_s^{P} \in \mathbf{R}^{N_p \times C_p}$. By default, C_v, C_p and C_x are equal to avoid additional overhead for unifying fea-



(a) Pixel-point view cooperation for 3D side (b) Alternative to 2D side (c) Attention module

Fig. 2 Pixel-point view cooperation and interaction. Part (a) treats 3D points as queries and adapts to the 3D segmentation branch. Part (b) replaces the dashed part in (a) by treating valid pixels with calibrated correspondences as queries to adapt to the 2D side. Part (c) shows the attention module in part (a). (Colored figures are available in the online version at https://link.springer.com/journal/11633)

ture dimensions. We concatenate ($\circ$) two of them in homo-modal and hetero-modal forms respectively for subsequent fusion via an MLP layer. After that, both forms are independently weighted ($\otimes$) by the learnable parameters $W_\circ$, and then another MLP layer is utilized to generate compact geometric point features. The final homo-modal view-wise cooperative features $\mathbf{F}_s^{\widetilde{VP}}$ are aggregated in a residual manner as follows:

$$\begin{aligned}\mathbf{F}_{s,g}^{\widetilde{VP}} &= \mathrm{ReLU}\left(\mathrm{MLP}_{s,g}^{\widetilde{VP}}(\mathbf{F}_s^{\tilde{V}} \circ \mathbf{F}_s^{\tilde{P}})\right)\\ \mathbf{F}_s^{\widetilde{VP}} &= \mathbf{F}_{s,g}^{\widetilde{VP}} \oplus \mathrm{ReLU}\left(\mathrm{MLP}_{s,l}^{\widetilde{VP}}(W_{s,\circ}^{\widetilde{VP}} \otimes \mathbf{F}_{s,g}^{\widetilde{VP}})\right)\end{aligned} \tag{3}$$

where $\mathbf{F}_s^{\widetilde{XP}} \in \mathbf{R}^{N_p \times C_x}$ can be generated by replacing the relevant variables from $\widetilde{VP}$ to $\widetilde{XP}$ in the process of generating $\mathbf{F}_s^{\widetilde{VP}} \in \mathbf{R}^{N_p \times C_v}$ but following the same (3).

Cross-view cooperative completion. Realistic multimodal samples may not always provide a comprehensive view, and unfortunately these incomplete data have to be discarded. In solving this issue, we introduce a cross-modal feature completion module to transform features that mimic each other, thereby enabling the learning of replaceable pseudo-modal features. Specifically, an MLP layer is first employed to map point features $\mathbf{F}_s^P$ to point-wise pseudo-pixel features $\mathbf{F}_s^{\widetilde{X}'}$, while another additional symmetric MLP layer is utilized to perform the conversion from point-wise pixel features $\mathbf{F}_s^{\widetilde{X}}$ to pseudo-point features $\mathbf{F}_s^{P'}$. Note that the gradient of the reference feature is not optimized to avoid contaminating local modal features, and points not covered by the multi-camera FOV are also ignored through the mentioned binary mask $\mathcal{M}$. Finally, a cooperative imitation constraint (i.e., L1 regularization loss $\mathcal{L}_{L1}$) is imposed between the simulated pseudo-modal features and their corresponding actual reference features as follows:

$$\begin{aligned}\mathcal{L}_{P\sim P'} &= \mathcal{L}_{L1}\left(\mathcal{M} \times (\mathbf{F}_s^P, \mathbf{F}_s^{P'})\right)\\ \mathcal{L}_{\widetilde{X}\sim\widetilde{X}'} &= \mathcal{L}_{L1}\left(\mathcal{M} \times (\mathbf{F}_s^{\widetilde{X}}, \mathbf{F}_s^{\widetilde{X}'})\right).\end{aligned} \tag{4}$$

3.4 Cross-view cooperative supervision

Point-view supervision. With the given 3D semantic labels $\mathbf{Y}^P$ with the number of categories N_C, a point-specific head $\mathcal{H}^P$ is applied to estimate point-wise segmentation results $\widehat{\mathbf{Y}}^P$ in (5), and optimized jointly using the cross-entropy loss $\mathcal{L}_{ce}$ and the Lovasz-softmax loss $\mathcal{L}_{lovasz}$:

$$\mathcal{L}_P = \mathcal{L}_{ce}\left(\widehat{\mathbf{Y}}^P, \mathbf{Y}^P\right) + \mathcal{L}_{lovasz}\left(\widehat{\mathbf{Y}}^P, \mathbf{Y}^P\right). \tag{5}$$

Voxel-view supervision. We extend supervision from the point-view to the voxel-view and similarly build a voxel-specific head $\mathcal{H}^V$ to predict voxel-wise segmentation results $\widehat{\mathbf{Y}}^V$ in (6). The voxel supervision labels $\mathbf{Y}^V$ are voxelized based on the point-wise labels $\mathbf{Y}^P$. The same voxel is considered as an ignored category (set to zero) once it contains points of more than one semantic category.

$$\mathcal{L}_V = \mathcal{L}_{ce}\left(\widehat{\mathbf{Y}}^V, \mathbf{Y}^V\right) + \mathcal{L}_{lovasz}\left(\widehat{\mathbf{Y}}^V, \mathbf{Y}^V\right). \tag{6}$$

Pixel-view supervision. The pixel-specific head $\mathcal{H}^X$ produces image-based 2D segmentation results $\widehat{\mathbf{Y}}^X$ in (7). In the absence of accurate 2D semantic annotation, each point is projected onto the image and rounded to the nearest pixel label $\mathbf{Y}^X$. Although the projected 2D semantics are sparse, it is still sufficient to support training and reasoning on the 2D side. Considering the computational burden, the step of upsampling to the original image size can be omitted. At this time, the annotations are downsampled to $\mathbf{Y}_i^X\left[c_i, \langle r^W u_i\rangle, \langle r^H v_i\rangle\right] = \mathbf{Y}_i^P$ according to the resolutions r^H and r^W but only suffer small accuracy fluctuations. The final sparse 2D semantic constraints are formalized as

$$\mathcal{L}_X = \mathcal{L}_{\mathrm{ce}}\left(\widehat{\mathbf{Y}}^X, \mathbf{Y}^X\right). \tag{7}$$

Completion-view supervision. Regardless of whether modal missing occurs in 2D-assisted 3D semantics or 3D-assisted 2D semantics, multimodal and cross-view feature completion can always be achieved with the help of the modules and constraints mentioned in Section 3.3. Ultimately, they work together to form a cooperative imitation loss $\mathcal{L}_C$ as follows:

$$\mathcal{L}_C = \mathcal{L}_{P\sim P'} + \mathcal{L}_{\widetilde{X}\sim\widetilde{X}'}. \tag{8}$$

Overall objective. Although each loss function is generic and common, there are still significant improvements after joint optimization in our multimodal cross-view segmentation workflow, i.e.,

$$\mathcal{L}_{overall} = \mathcal{L}_P + \alpha\mathcal{L}_V + \beta\mathcal{L}_X + \gamma\mathcal{L}_C. \tag{9}$$

where α, β, and γ are the loss coefficients to balance the effect of each loss term.

3.5 Multimodal pre-synchronization

Timestamp pre-synchronization. If the LiDAR sensor and the multi-camera sensors are not guaranteed to be employed at the same frequency, generating trainable multimodal sample pairs requires cross-modal frame synchronization. Specifically, the LiDAR frame with timestamp t_l in the global coordinate system is aligned to the camera frame with timestamp t_c as follows:

$$T_{cam\leftarrow LiDAR} = T_{cam\leftarrow ego(t_c)} \times T_{ego(t_c)\leftarrow glo} \times T_{glo\leftarrow ego(t_l)} \times T_{ego(t_l)\leftarrow LiDAR}. \quad (10)$$

Labeling pre-synchronization. Although benchmarks for LiDAR-based and image-based segmentation have been established separately and extensively studied, realizing multimodal simultaneous annotation for full-volume data still faces difficulties. Taking the Waymo dataset[49] as an example, the number of valid 3D annotations in the training set is 23 691, and the number of valid 2D annotations is 12 295, however, only 1 852 samples coexist with both 2D and 3D annotations. In addition, the semantic categories in the 2D and 3D directions are not identical. To tackle this issue, we propose a targeted pre-training and pre-synchronization strategy. In essence, we train the two modality segmentation tasks independently using the underlying backbone to fully absorb modality-specific semantic information. Subsequently, 3D semantic projection and 2D semantic lifting are implemented by camera calibration to filter out the union of annotated sets for fine-tuning. For the image-based branch, we need to restore its original decoding head and further fine-tune it to complete the entire training process.

Augmentation pre-synchronization. In our setup, we pre-map the point-pixel correspondence, while the point-voxel correspondence is locally interpolated. Preserving the order of the voxel-point-pixel chain of each point can avoid semantic ambiguity caused by modality-specific augmentations. This allows us to adhere to the common augmentations that each modality uses independently within multimodal augmentations in a straightforward manner. Specifically, as long as they do not alter the image and pixel content, various affine transformations are deemed acceptable for both the point-view and the pixel-view input, including but not limited to global rotation, global translation, and global scaling. Multi-camera data augmentation can be further decomposed into independent sub-operations for each viewpoint, encompassing but not limited to random scaling, random cropping, random flipping, and color jittering. During the inference process, we enable test-time augmentation through voting. This involves rotating 12 views of the input scene along the Z-axis and averaging the predictions. For inference, we enable test-time augmentation with voting, i.e., rotate 12 views of the input scene along the Z-axis and average the predictions. In short, pre-synchronized augmentation for LiDAR and multi-camera can maintain multimodal semantic consistency while taking into account their common augmentation operators. And this flexibility provides a more general data interface for multimodal semantic segmentation.

4 Experiments

Datasets. We conduct experiments on two popular benchmarks, SemanticKITTI[50] dataset and Waymo Open[49] dataset. The SemanticKITTI dataset is collected by a 64-beams LiDAR system and composed of 22 point cloud sequences. Following [12, 37], sequences from 00 to 10 are divided into training sets with 19 130 samples and sequence 08 is used for validation set with 4 071 samples. The remaining sequences 11 to 21 are adopted as the testing set. It covers common scenarios in driving scenarios, with a total of 20 categories recorded as N_C (including one undefined category), and only provides images from the front-view camera. The Waymo Open dataset (WOD) consists of 23 691 training samples, 5 976 validation samples, and 2 982 test samples[49]. Each sample contains an RGB image with 64 LiDAR frames and captured by 5 cameras: front, front left, front right, side left, and side right, while the rear view is not provided. Both the first and second return point clouds need to be segmented. The total number of categories N_C is 23, including one ignored and 22 valid semantic categories. Note that not all frames provide 2D and 3D segmentation labels, so we adopt the pre-annotation scheme mentioned in Section 3.5.

Evaluation metric. The mean intersection-over-union (mIoU) is used as the evaluation metric, defined as $\frac{1}{C}\sum_{c=1}^{N_C-1}\frac{TP_c}{TP_c+FP_c+FN_c}$, where TP_c, FP_c, FN_c denote the counts of true positive, false positive, and false negative predictions for the c-th category among the N_C-1 valid categories.

Multimodal segmenter. The powerful point voxel backbone Minkowski-UNet34[51] is chosen as the segmenter for the point cloud branches, and its entire structure is based on UNet, which facilitates the insertion of our cross-view cooperative pathway module at the end of each stage. In order to take into account both lightweight and high performance, SegFormer-B2[8] serves as the segmenter of the image branch. Other alternative backbones are given in Table 1 and demonstrate the scalability of our CoSEG pipeline.

Table 1 Scalability analysis by barely varying the backbone setting on Waymo *val* set

Backbone type	mIoU	#Mem(GB)	InfTime(fps)
SegFormer-B0[8]	45.58	3.60	30.88
ResNet-101[52]	39.35	3.90	10.29
ConvNeXt-B[46]	52.13	8.50	10.88

Implementation details. For data processing, we perform voxelization of point clouds within a specific range of each dataset and simultaneously resize the images. For SemanticKITTI, the voxelization range is from [−75.2 m, −75.2 m, −4.0 m] to [+75.2 m, +75.2 m, +2.0 m], while the image is resized to $360 \times 1\,280$. For Waymo Open, the voxelization range is from [−75.2 m, −75.2 m, −2.0 m] to [+75.2 m, +75.2 m, +4.0 m], while the image size of each multi-camera view is adjusted to

640×960. The voxel size always remains (0.1 m, 0.1 m, 0.15 m). For pre-training of the workflow, all models first follow the original training configuration[8, 51] of the individual modalities. Then update to the unified fine-tuning configuration, that is, enable the AdamW optimizer, and the polynomial scheduler with a division factor of 10 and a maximum learning rate of 0.01. The batch size is set to 8 to adapt to the throughput of multimodal samples, and the fine-tuning process lasts for 12 epochs. All the experiments are arranged on NVIDIA RTX A6 000 GPUs. It is worth noting that several LiDAR-only baseline models require larger batch size and more iterations to achieve relatively better results compared to the described training setup.

Data augmentation operators. We apply different data augmentation operators to the LiDAR point and camera image branches. Following [8], we apply data augmentation to the 2D image input via random resizing, random horizontal flipping and random cropping with a ratio of 0.5–2.0. For the point cloud branch, we perform random flipping along the X-axis or Y-axis, followed by random translation up to a maximum distance of 0.1, global scaling by random factors in the range $[0.9, 1.1]$, and global rotation by random angles in the range $[0, 2\pi]$. Additionally, in the pre-training phase of the 3D branch, LaserMix[53] and PolarMix[54] are incorporated, but they are excluded during multimodal cooperative fine-tuning. The rationale behind this decision lies in the validated effectiveness of directly mixing point clouds, whereas directly mixing sparsely sampled pixels is prone to introduce noise interference.

Inference. In the inference stage, the argmax function is applied to the point-wise 3D output $\widehat{\mathbf{Y}}^P$ and pixel-wise 2D output $\widehat{\mathbf{Y}}^X$, so that the category indexes with the highest scores are regarded as the segmentation results. To prepare the results for submission, we employ test-time augmentation as a common practice for other submissions without the tricks of model ensembles.

4.1 Comparative study

Quantitative comparison with SOTA methods. We provide a relatively comprehensive comparison between CoSEG and competing LiDAR segmentation networks. Table 2 shows the class-wise IoU scores of different LiDAR semantic segmentation methods on the SemanticKITTI test set. Among all LiDAR segmentation algorithms, CoSEG achieves significant performance gains with its multimodal cross-view cooperative aggregation and completion, surpassing the competing 2DPASS[42] by 1.6 mIoU, especially for various vehicle categories. Table 3 shows the class-wise IoU scores of different LiDAR semantic segmentation methods on the Waymo Open value set. We can observe that CoSEG also outperforms the other solutions, i.e., achieves a higher efficacy of 2.4 mIoU over our implemented SPVCNN[11], and similarly a gain of 1.3 mIoU over the baseline that optimizes the LiDAR branches individually using multimodal inputs. Note that the comparison scenarios all use the same training configuration for a fair comparison. The impressive results shown in both benchmarks illustrate that our approach not only outperforms the LiDAR-only approach, but also has significant advantages over other multimodal solutions. In addition, we provide the class-wise IoU scores for camera-based 2D semantic segmentation on the Waymo Open value set in Table 4. Jointly optimizing the multimodal input in CoSEG for 2D segmentation tasks still yields a 1.5 mIoU improvement compared to the baseline with only multi-camera images. This once again underscores the mutually beneficial nature of multimodal cross-view cooperation.

4.2 Ablation study

Ablation for augmentation pre-synchronization. The augmentation pre-synchronization creates the conditions for multimodal inputs to simultaneously perform modality-specific augmentation without breaking geometric and semantic consistency. We verify the effectiveness of this module by gradually increasing the LiDAR-only augmentation A_{LiDAR}, the camera-only augmentation A_{cam} and the multimodal joint pre-synchronized augmentation A_{syn} in Table 5. mIoU refers to the evaluation metric for 360° point clouds, while mIoU* means that only point clouds within the overlap of multi-camera FOV are evaluated, as reported in [38]. Note that the cross-view cooperative interaction module is unable to rescue points outside the FOV, so mIoU is relatively lower than mIoU*. Nevertheless, the pre-synchronization strategy for augmentation effectively minimizes the disparity between mIoU and mIoU* from 3.7 to 0.7 by integrating a range of diverse multimodal samples.

Ablation for the cross-view cooperative mechanism. Table 6 shows the impacts of cross-view cooperative interaction modules on multimodal segmentation. It can be seen that the interactive capability of cross-view pathways improves the mIoU* metric from 67.6 to 68.7, but has less effect on mIoU*, proving that the multimodal cross-view features are complementary and mutually beneficial. On the other hand, the complementary ability of cross-view paths alleviates the dilemma of missing features at the external point of the multi-camera FOV, and it narrows the gap between mIoU and mIoU*. This demonstrates that the cross-view cooperative completion module improves global reasoning capabilities by imitating the generation of pseudo-features, promoting the joint optimization of multimodal features instead of falling into the local optimum of individual modalities.

Robust ablation for the camera malfunction. In Table 7, the CoSEG workflow shows strong performance in challenging scenarios (e.g., multi-camera partial failures) and circumvents the risk of model crashes when

Table 2 Performance comparison of CoSEG and SoTA LiDAR-based semantic segmentation methods on the SemanticKITTI *test* set

Methods	mIoU	Car	Bicy	Moto	Truc	O.veh	Ped	B.list	M.list	Road	Park	Walk	O.gro	Build	Fenc	Veg	Trun	Terr	Pole	Sign
AMVNet[32]	65.3	96.2	59.9	54.2	48.8	45.7	71.0	65.7	11.0	90.1	71.0	75.8	32.4	92.4	69.1	85.6	71.7	69.6	62.7	67.2
PolarNet[34]	54.3	90.8	**74.4**	61.7	21.7	**90.0**	**93.8**	22.9	40.3	30.1	28.5	**84.0**	**65.5**	67.8	43.2	40.2	5.6	61.3	51.8	57.5
SPVNAS[11]	66.4	97.3	51.5	50.8	59.8	58.8	65.7	65.2	43.7	90.2	67.6	75.2	16.9	91.3	65.9	86.1	73.4	71.0	64.2	66.9
Cylinder3D[55]	68.9	97.1	67.6	63.8	50.8	58.5	73.7	69.2	48.0	92.2	65.0	77.0	32.3	90.7	66.5	85.6	72.5	69.8	62.4	66.2
RPVNet[37]	70.3	**97.6**	68.4	68.7	44.2	61.1	75.9	74.4	73.4	**93.4**	70.3	80.7	33.3	**93.5**	72.1	86.5	75.1	71.7	64.8	61.4
2DPASS[42]	72.9	97.0	63.6	63.4	61.1	61.5	77.9	**81.3**	**74.1**	89.7	67.4	74.7	40.0	**93.5**	**72.9**	86.2	73.9	71.0	65.0	**70.4**
CoSEG (ours)	**74.5**	97.1	71.2	**74.4**	**62.9**	73.4	78.0	74.2	59.7	91.8	**73.4**	78.6	45.5	92.8	71.8	**86.9**	**75.9**	**72.5**	**67.6**	68.6

Table 3 Performance comparison of CoSEG and SoTA LiDAR-based semantic segmentation methods on the *val* set of Waymo Open dataset. † indicates that the method is our implementation. ◊ indicates that the LiDAR branches are individually optimized using multimodal inputs.

Methods	mIoU	Car	Truck	Bus	Other vehicle	Motor-cyclist	Bicyclist	Pedestrian	Sign	Traffic light	Pole	Cons. cone	Bicycle	Motor-cycle	Building	Vegetation	Tree trunk	Curb	Road	Lane marker	Other ground	Walkable	Sidewalk
Cylinder3D†[55]	65.5	94.1	59.2	73.8	27.8	**2.5**	60.9	86.8	70.5	34.3	74.0	64.4	61.2	76.0	94.2	90.1	66.0	63.9	**93.3**	48.6	45.5	78.9	75.0
SPVCNN†[11]	66.3	92.8	57.5	77.5	25.6	0.0	**70.4**	87.1	74.1	38.1	74.8	68.1	70.3	79.1	92.4	89.9	64.5	65.7	90.4	50.3	41.7	76.7	73.1
CoSEG◊ (ours)	67.4	93.3	57.5	77.9	37.8	0.3	69.9	87.0	72.3	40.1	73.0	69.1	**72.7**	78.6	94.6	88.3	67.7	65.4	90.5	**52.7**	46.2	77.9	72.6
CoSEG (ours)	**68.7**	**94.8**	**61.2**	**85.4**	**39.2**	1.4	69.6	**89.8**	**75.1**	**41.1**	**76.5**	**69.5**	69.2	**80.4**	**95.8**	**91.8**	**68.3**	**70.0**	93.0	48.8	**47.3**	**79.6**	**77.0**

Table 4 Performance comparison of CoSEG and SoTA image-based semantic segmentation methods on the *val* set of Waymo Open dataset. L means LiDAR-wise input is provided, C means camera-wise input is provided.

L	C	mIoU	Undefined	Car	Truck	Bus	Other vehicle	Motor-cyclist	Bicyclist	Pedestrian	Sign	Traffic light	Pole	Cons. cone	Bicycle	Motor-cycle	Building	Vegetation	Tree trunk	Curb	Road	Lane marker	Other ground	Walkable	Sidewalk
–	✓	49.8	9.6	77.8	48.3	59.7	28.5	2.0	35.8	64.1	48.6	31.0	54.9	42.5	40.9	47.5	80.3	75.8	50.9	48.5	85.0	37.5	40.6	69.1	66.6
✓	✓	51.3	10.3	79.0	52.1	68.1	31.5	0.4	38.2	64.7	48.9	30.4	55.8	42.9	40.0	57.1	80.7	75.9	51.2	49.4	85.5	39.4	41.2	69.9	67.6

Table 5 Analysis of the ablation performance of LiDAR-based semantic segmentation for pre-synchronized augmentation on the Waymo Open dataset *val* set. Ablation is performed by gradually adding LiDAR-based augmentation A_{LiDAR}, camera-based augmentation A_{cam} and pre-synchronized augmentation A_{syn}.

A_{LiDAR}	A_{cam}	A_{syn}	mIoU	mIoU*
			61.6	65.3
✓			67.4	67.6
✓	✓		65.9	66.9
✓	✓	✓	68.8	69.5

Table 6 Ablation for the cross-view cooperative interaction and completion on the Waymo Open dataset *val* set. MCI denotes multimodal cross-view cooperative interaction and MCC denotes multimodal cross-view cooperative completion. The gap can be formulated as mIoU-mIoU*.

MCI	MCC	mIoU	mIoU*	Gap
		67.4	67.6	-0.2
✓		67.1	68.7	+1.6
✓	✓	68.8	69.5	-0.7

Table 7 Robustness analysis by removing camera views on the Waymo Open dataset *val* set. #Cam represents the number of views available for the multi-camera and × represents the LiDAR-only baseline.

#Cam	×	1	3	5
–	67.4	67.9	68.0	68.8

processing partially visible scenes. Notably, CoSEG outperforms the LiDAR-only baseline even without all camera views, highlighting the effectiveness of our interactive completion mechanism in facilitating cross-modal information transfer. This efficiency is attributed to the consistent imitation constraint $\mathcal{L}_C$, which uses priors on image appearance to guide point cloud grouping, while taking advantage of point cloud object-oriented clustering to guide pixel-level semantics distinguish.

Analysis for multi-frame superimposing. Although single-frame point clouds are relatively sparse, the spatial localisation of adjacent frames is close, so multi-frame superposition of point clouds is a commonly used technique for LiDAR-based 3D perception tasks. Following [56], we collapse multiple previous frames into the current frame based on the vehicle′s ego-motion. As shown in Table 8, multi-frame point cloud superposition is beneficial to object segmentation, especially for some static objects (such as poles, cones, signs, sidewalks, roads, etc.) or slowly moving objects (such as pedestrians). This is because the relative motion between a high-speed moving object and the vehicle cannot be accurately estimated, resulting in motion blur. However, excessive point cloud frame stacking should be avoided as the semantic ambiguity caused by motion blur can have side effects.

Table 8 Ablation for multiple LiDAR frame superimposing on the Waymo Open dataset *val* set. #LiDAR is the frame number of LiDAR. #Camera is whether use camera-views or not.

#LiDAR	1	5	10	15	20	10
#Cam	×	×	×	×	×	✓
mIoU	68.8	69.7	71.1	69.8	67.5	71.5

Complexity scalability. Table 8 shows the difference in performance gains provided by different backbones in multimodal segmentation tasks. Due to the conflict between multimodal input and limited computing resources, we adopt the scalable SegFormer[8] and apply its B2 variant as the default backbone in the image branch. The novel network[57] can further accelerate workflow efficiency. Overall, CoSEG can be arbitrarily replaced with various backbones based on complexity and scalability, allowing for a flexible trade-off between performance and efficiency.

4.3 Qualitative results

As illustrated in Fig. 3, the proposed multimodal cooperative segmentation workflow (CoSEG) can simultaneously obtain pixel-level and point-level segmentation results. Due to the insufficient alignment and completeness of multimodal semantic annotations in the SemanticKITTI and Waymo Open datasets, we rely on camera calibration projections to obtain a few but reliable pixel-level sparse semantic labels. However, it should be noted that the projection does not cover the entire camera acquisition range, resulting in errors in some 2D semantics, such as the region of sky, and it has not been well addressed in previous multimodal segmentation methods. Our multimodal cooperative segmentation can fully include various supervised or unsupervised data for flexible training, thus greatly improving the scalability of the model.

5 Conclusions

The proposed CoSEG workflow, a multimodal cooperative segmentation approach, is designed to concurrently conduct LiDAR semantic segmentation and image semantic segmentation, facilitating cross-view interaction and completion in a multimodal context. Our method adeptly addresses the challenge of heterogeneous multimodal feature fusion by establishing voxel-point-pixel cross-view feature pathways centered around point data. Additionally, pre-synchronization of multimodal data serves as a gateway for various data enhancements, contributing to the enhanced robustness of multimodal segmentation. Extensive experiments conducted on the SemanticKITTI and Waymo datasets validate that CoSEG attains significant improvements and attains state-of-the-art performance. Looking ahead, we aspire to delve more deeply into

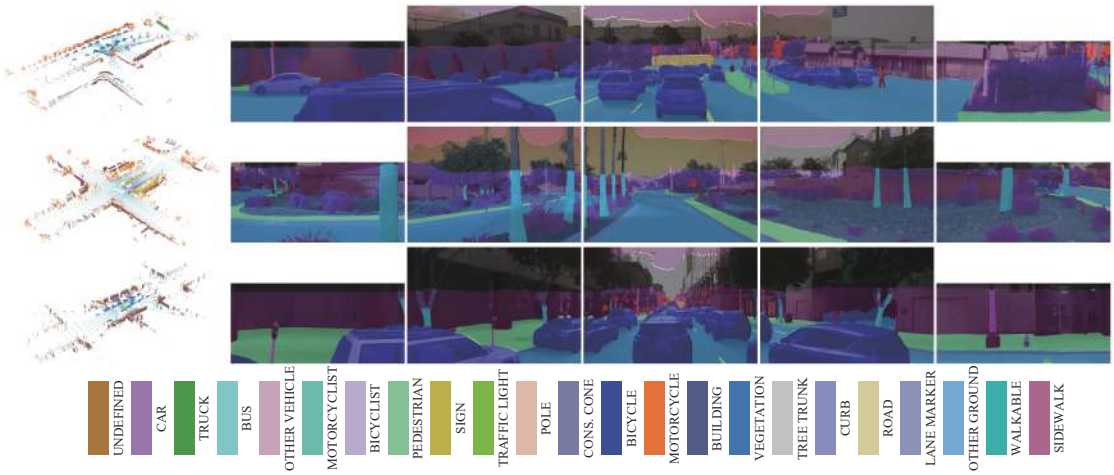


Fig. 3 Qualitative results on Waymo *val* set. Best viewed in color and by zoom-in. Due to a lack of point projection, the upper half of the image is not annotated. Each row from left to right: LiDAR, CAM_FRONT, CAM_FRONT_LEFT, CAM_FRONT_RIGHT, CAM_SIDE_LEFT and CAM_SIDE_RIGHT. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

the potential of multimodal semantic segmentation by exploring the capabilities of large language models.

Acknowledgements

This work was supported in part by the National Key R&D Program of China (No. 2022ZD0160102), the National Natural Science Foundation of China (Nos. 61836014, U21B2042, 62072457 and 62006231).

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

References

[1] B. Gao, Y. C. Pan, C. K. Li, S. B. Geng, H. J. Zhao. Are we hungry for 3D LiDAR data for semantic segmentation? A survey of datasets and methods. *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6063–6081, 2022. DOI: 10.1109/TITS.2021.3076844.

[2] G. Rizzoli, F. Barbato, P. Zanuttigh. Multimodal semantic segmentation in autonomous driving: A review of current approaches and future perspectives. *Technologies*, vol. 10, no. 4, Article number 90, 2022. DOI: 10.3390/technologies10040090.

[3] K. L. Huang, B. T. Shi, X. Li, X. Li, S. Y. Huang, Y. K. Li. Multi-modal sensor fusion for auto driving perception: A survey, [Online], Available: http://arxiv.org/abs/2202.02703, 2024.

[4] C. Peng, X. Y. Zhang, G. Yu, G. M. Luo, J. Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 1743–1751, 2017. DOI: 10.1109/CVPR.2017.189.

[5] H. S. Zhao, J. P. Shi, X. J. Qi, X. G. Wang, J. Y. Jia. Pyramid scene parsing network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 6230–6239, 2017. DOI: 10.1109/CVPR.2017.660.

[6] Z. L. Huang, X. G. Wang, L. C. Huang, C. Huang, Y. C. Wei, W. Y. Liu. CCNet: Criss-cross attention for semantic segmentation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, pp. 603–612, 2019. DOI: 10.1109/ICCV.2019.00069.

[7] C. X. Liu, L. C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, L. Fei-Fei. Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, pp. 82–92, 2019. DOI: 10.1109/CVPR.2019.00017.

[8] E. Z. Xie, W. H. Wang, Z. D. Yu, A. Anandkumar, J. M. Álvarez, P. Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 12077–12090, 2021.

[9] W. Q. Zhang, Z. L. Huang, G. Z. Luo, T. Chen, X. G. Wang, W. Y. Liu, G. Yu, C. H. Shen. TopFormer: Token pyramid transformer for mobile semantic segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 12073–12083, 2022. DOI: 10.1109/CVPR52688.2022.01177.

[10] Q. Y. Hu, B. Yang, L. H. Xie, S. Rosa, Y. L. Guo, Z. H. Wang, N. Trigoni, A. Markham. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 11105–11114, 2020. DOI: 10.1109/CVPR42600.2020.01112.

[11] H. T. Tang, Z. J. Liu, S. Y. Zhao, Y. J. Lin, J. Lin, H. R. Wang, S. Han. Searching efficient 3D architectures with sparse point-voxel convolution. In *Proceedings of the 16th European Conference on Computer Vision*, Glasgow, UK, pp. 685–702, 2020. DOI: 10.1007/978-3-030-58604-1_41.

[12] X. G. Zhu, H. Zhou, T. Wang, F. Z. Hong, Y. X. Ma, W. Li, H. S. Li, D. H. Lin. Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Nashville, USA, pp. 9939–9948, 2021.

[13] T. Cortinhal, G. Tzelepis, E. Erdal Aksoy. SalsaNext: Fast, uncertainty-aware semantic segmentation of LiDAR point clouds. In *Proceedings of the 15th International Symposium Advances in Visual Computing*, San Diego, USA, pp. 207–222, 2020. DOI: 10.1007/978-3-030-64559-5_16.

[14] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018. DOI: 10.1109/TPAMI.2017.2699184.

[15] L. C. Chen, Y. K. Zhu, G. Papandreou, F. Schroff, H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp. 801–818, 2018. DOI: 10.1007/978-3-030-01234-2_49.

[16] C. Q. Yu, J. B. Wang, C. X. Gao, G. Yu, C. H. Shen, N. Sang. Context prior for scene segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 12413–12422, 2020. DOI: 10.1109/CVPR42600.2020.01243.

[17] H. Zhang, K. Dana, J. P. Shi, Z. Y. Zhang, X. G. Wang, A. Tyagi, A. Agrawal. Context encoding for semantic segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 7151–7160, 2018. DOI: 10.1109/CVPR.2018.00747.

[18] M. M. Zhen, J. L. Wang, L. Zhou, S. W. Li, T. W. Shen, J. X. Shang, T. Fang, L. Quan. Joint semantic segmentation and boundary detection using iterative pyramid contexts. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 13663–13672, 2020. DOI: 10.1109/CVPR42600.2020.01368.

[19] H. H. Ding, X. D. Jiang, A. Q. Liu, N. M. Thalmann, G. Wang. Boundary- aware feature propagation for scene segmentation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, pp. 6818–6828, 2019. DOI: 10.1109/ICCV.2019.00692.

[20] A. Shaw, D. Hunter, F. Landola, S. Sidhu. SqueezeNAS: Fast neural architecture search for faster semantic segmentation. In *Proceedings of IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, Republic of Korea, pp. 2014–2024, 2019. DOI: 10.1109/ICCVW.2019.00251.

[21] J. Jain, J. C. Li, M. T. Chiu, A. Hassani, N. Orlov, H. Shi. OneFormer: One transformer to rule universal image segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 2989–2998, 2023. DOI: 10.1109/CVPR52729.2023.00292.

[22] R. Q. Charles, H. Su, M. Kaichun, L. J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 77–85, 2017. DOI: 10.1109/CVPR.2017.16.

[23] R. Q. Charles, L. Yi, H. Su, L. J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, USA, pp. 5105–5114, 2017.

[24] Y. Wang, Y. B. Sun, Z. W. Liu, S. E. Sarma, M. M. Bronstein, J. M. Solomon. Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics*, vol. 38, no. 5, Article number 146, 2019. DOI: 10.1145/3326362.

[25] W. X. Wu, Z. G. Qi, F. X. Li. PointConv: Deep convolutional networks on 3D point clouds. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, pp. 9613–9622, 2019. DOI: 10.1109/CVPR.2019.00985.

[26] H. Thomas, C. R. Qi, J. E. Deschaud, B. Marcotegui, F. Goulette, L. J. Guibas. KPConv: Flexible and deformable convolution for point clouds. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, pp. 6410–6419, 2019. DOI: 10.1109/ICCV.2019.00651.

[27] B. S. Hua, M. K. Tran, S. K. Yeung. Pointwise convolutional neural networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 984–993, 2018. DOI: 10.1109/CVPR.2018.00109.

[28] F. J. Lawin, M. Danelljan, P. Tosteberg, G. Bhat, F. S. Khan, M. Felsberg. Deep projective 3D semantic segmentation. In *Proceedings of the 17th International Conference on Computer Analysis of Images and Patterns*, Ystad, Sweden, pp. 95–107, 2017. DOI: 10.1007/978-3-319-64689-3_8.

[29] M. Tatarchenko, J. Park, V. Koltun, Q. Y. Zhou. Tangent convolutions for dense prediction in 3D. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 3887–3896, 2018. DOI: 10.1109/CVPR.2018.00409.

[30] B. C. Wu, A. Wan, X. Y. Yue, K. Keutzer. SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud. In *Proceedings of IEEE International Conference on Robotics and Automation*, Brisbane, Australia, pp. 1887–1893, 2018. DOI: 10.1109/ICRA.2018.8462926.

[31] B. C. Wu, X. Y. Zhou, S. C. Zhao, X. Y. Yue, K. Keutzer. SqueezeSegV2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud. In *Proceedings of International Conference on Robotics and Automation*, Montreal, Canada, pp. 4376–4382, 2019. DOI: 10.1109/ICRA.2019.8793495.

[32] V. E. Liong, T. H. T. Nguyen, S. Widjaja, D. Sharma, Z. J. Chong. AMVNet: Assertion-based multi-view fusion network for LiDAR semantic segmentation, [Online], Available: https://arxiv.org/abs/2012.04934, 2023.

[33] B. Graham, M. Engelcke, L. Van Der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 9224–9232, 2018. DOI: 10.1109/CVPR.2018.00961.

[34] Y. Zhang, Z. X. Zhou, P. David, X. Y. Yue, Z. R. Xi, B. Q. Gong, H. Foroosh. PolarNet: An improved grid representation for online LiDAR point clouds semantic segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 9598–9607, 2020. DOI: 10.1109/CVPR42600.2020.00962.

[35] X. Lai, Y. K. Chen, F. B. Lu, J. H. Liu, J. Y. Jia. Spherical transformer for LiDAR-based 3D recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 17545–17555, 2023. DOI: 10.1109/CVPR52729.2023.01683.

[36] R. Cheng, R. Razani, Y. Ren, B. B. Liu. S3Net: 3D LiDAR sparse semantic segmentation network. In *Proceedings of IEEE International Conference on Robotics and Automation*, Xi'an, China, pp. 14040–14046, 2021. DOI: 10.1109/ICRA48506.2021.9561305.

[37] J. Y. Xu, R. X. Zhang, J. Dou, Y. S. Zhu, J. Sun, S. L. Pu. RPVNet: A deep and efficient range-point-voxel fusion network for LiDAR point cloud segmentation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, pp. 16004–16013, 2021. DOI: 10.1109/ICCV48922.2021.01572.

[38] Z. W. Zhuang, R. Li, K. Jia, Q. C. Wang, Y. Q. Li, M. K. Tan. Perception-aware multi-sensor fusion for 3D LiDAR semantic segmentation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, pp. 16260–16270, 2021. DOI: 10.1109/ICCV48922.2021.01597.

[39] K. El Madawi, H. Rashed, A. El Sallab, O. Nasr, H. Kamel, S. Yogamani. RGB and LiDAR fusion based 3D semantic segmentation for autonomous driving. In *IEEE Intelligent Transportation Systems Conference*, Auckland, New Zealand, pp. 7–12, 2019. DOI: 10.1109/ITSC.2019.8917447.

[40] G. Krispel, M. Opitz, G. Waltner, H. Possegger, H. Bischof. FuseSeg: LiDAR point cloud segmentation fusing multi-modal data. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, Snowmass, USA, pp. 1863–1872, 2020. DOI: 10.1109/WACV45572.2020.9093584.

[41] S. Vora, A. H. Lang, B. Helou, O. Beijbom. PointPainting: Sequential fusion for 3d object detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 4603–4611. DOI: 10.1109/CVPR42600.2020.00466.

[42] X. Yan, J. T. Gao, C. D. Zheng, C. Zheng, R. M. Zhang, S. G. Cui, Z. Li. 2DPASS: 2D priors assisted semantic segmentation on LiDAR point clouds. In *Proceedings of the 17th European Conference on Computer Vision*, Tel Aviv-Yafo, Israel, pp. 677–695, 2022. DOI: 10.1007/978-3-031-19815-1_39.

[43] C. W. Wang, C. Ma, M. Zhu, X. K. Yang. PointAugmenting: Cross-modal augmentation for 3D object detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, pp. 11789–11798, 2021. DOI: 10.1109/CVPR46437.2021.01162.

[44] X. P. Wu, L. Peng, H. H. Yang, L. Xie, C. X. Huang, C. Q. Deng, H. F. Liu, D. Cai. Sparse fuse dense: Towards high quality 3D detection with depth completion. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 5408–5417, 2022. DOI: 10.1109/CVPR52688.2022.00534.

[45] X. Y. Bai, Z. Y. Hu, X. G. Zhu, Q. Q. Huang, Y. L. Chen, H. B. Fu, C. L. Tai. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 1080–1089, 2022. DOI: 10.1109/CVPR52688.2022.00116.

[46] Z. Liu, H. Z. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, S. N. Xie. A convNet for the 2020s. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 11966–11976, 2022. DOI: 10.1109/CVPR52688.2022.01167.

[47] Y. Wang, V. Guizilini, T. Y. Zhang, Y. L. Wang, H. Zhao, J. Solomon. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. In *Proceedings of the 5th Conference on Robot Learning*, London, UK, pp. 180–191, 2021.

[48] H. T. Tang, Z. J. Liu, X. Y. Li, Y. J. Lin, S. Han. Torchsparse: Efficient point cloud inference engine. In *Proceedings of the 5th Machine Learning and Systems*, Santa Clara, USA, pp. 302–315, 2022.

[49] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. N. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. F. Chen, D. Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 2443–2451, 2020. DOI: 10.1109/CVPR42600.2020.00252.

[50] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, J. Gall. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, pp. 9296–9306, 2019. DOI: 10.1109/ICCV.2019.00939.

[51] C. Choy, J. Y. Gwak, S. Savarese. 4D spatio-temporal convNets: Minkowski convolutional neural networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, pp. 3070–3079, 2019. DOI: 10.1109/CVPR.2019.00319.

[52] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[53] L. D. Kong, J. W. Ren, L. Pan, Z. W. Liu. LaserMix for semi-supervised LiDAR semantic segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 21706–21716, 2023. DOI: 10.1109/CVPR52729.2023.02079.

[54] A. R. Xiao, J. X. Huang, D. Y. Guan, K. W. Cui, S. J. Lu, L. Shao. PolarMix: A general data augmentation technique for LiDAR point clouds. In *Proceedings of the 36th Conference on Neural Information Processing Systems*, New Orleans, USA, pp. 11035–11048, 2022.

[55] H. Zhou, X. G. Zhu, X. Song, Y. X. Ma, Z. Wang, H. S. Li, D. H. Lin. Cylinder3D: An effective 3D framework for driving-scene LiDAR semantic segmentation, [Online], Available: https://arxiv.org/abs/2008.01550, 2023.

[56] T. W. Yin, X. Y. Zhou, P. Krähenbühl. Center-based 3D object detection and tracking. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, pp. 11779–11788, 2021. DOI: 10.1109/CVPR46437.2021.01161.

[57] M. X. Tan, Q. V. Le. EfficientNetV2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning*, Vienna, Austria, pp. 10096–10106, 2021.

He Guan received the B. Sc. degree in automation from Harbin Institute of Technology, China in 2015, and the M. Sc. degree in computer science and technology from the Institute of Automation, Chinese Academy of Sciences, China in 2018. He is now a Ph. D. degree candidate with the University of Chinese Academy of Sciences, China.

His research interests include computer graphics and computer vision.

E-mail: sylcito@gmail.com

ORCID iD: 0000-0002-8827-5935

Chunfeng Song received the Ph. D. degree in pattern recognition and intelligent system from University of Chinese Academy of Sciences, China in 2020. He is now working at Shanghai Artificial Intelligence Laboratory, China. He has published more than 20 conference and journal papers such as IEEE TPAMI, TIP, IJCV, CVPR, ECCV, and AAAI.

His research interests include person identification, image segmentation, and unsupervised learning.

E-mail: chunfeng.song@nlpr.ia.ac.cn

ORCID iD: 0000-0003-1223-3242

Zhaoxiang Zhang received the B. Sc. degree in circuits and systems from the University of Science and Technology of China, China in 2004, and the Ph. D. degree in pattern recognition and intelligent system, Institute of Automation, Chinese Academy of Sciences, China in 2009. He is now a full professor in the New Laboratory of Pattern Recognition and the State Key Laboratory of Multimodal Artificial Intelligence Systems, China. Specifically, he recently focuses on biologically inspired intelligent computing and its applications in human analysis and scene understanding. He has published more than 150 papers in international journals and conferences, such as IEEE TPAMI, TIP, TIFS, IJCV, CVPR, ICCV, ECCV, and NeurIPS.

His research interests include computer vision, pattern recognition, and machine learning.

E-mail: zhaoxiang.zhang@ia.ac.cn (Corresponding author)

ORCID iD: 0000-0003-2648-3875

Machine Intelligence Research

Citation: H. Guan, C. Song, Z. Zhang. Lidar–camera cooperative semantic segmentation. *Machine Intelligence Research*, vol.22, no.5, pp.956–968, 2025. https://doi.org/10.1007/s11633–024–1508–2

Articles may interest you

Yolo-core: contour regression for efficient instance segmentation. *Machine Intelligence Research*, vol.20, no.5, pp.716-728, 2023.
DOI: 10.1007/s11633-022-1379-3

Edge-aware feature aggregation network for polyp segmentation. *Machine Intelligence Research*, vol.22, no.1, pp.101-116, 2025.
DOI: 10.1007/s11633-023-1479-8

Ecg biometrics via enhanced correlation and semantic-rich embedding. *Machine Intelligence Research*, vol.20, no.5, pp.697-706, 2023.
DOI: 10.1007/s11633-022-1345-0

Deep learning-based moving object segmentation: recent progress and research prospects. *Machine Intelligence Research*, vol.20, no.3, pp.335-369, 2023.
DOI: 10.1007/s11633-022-1378-4

Unveiling the hidden interactions among features: a heterogeneous graph approach for personality prediction. *Machine Intelligence Research*, vol.22, no.1, pp.91-100, 2025.
DOI: 10.1007/s11633-024-1495-3

Rethinking polyp segmentation from an out-of-distribution perspective. *Machine Intelligence Research*, vol.21, no.4, pp.631-639, 2024.
DOI: 10.1007/s11633-023-1472-2

Multitask learning with multiscale residual attention for brain tumor segmentation and classification. *Machine Intelligence Research*, vol.20, no.6, pp.897-908, 2023.
DOI: 10.1007/s11633-022-1392-6

WeChat: MIR

Twitter: MIR_Journal