2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

DefFusion: Deformable Multimodal Representation Fusion for 3D Semantic Segmentation

Rongtao Xu[1], Changwei Wang[1], Duzhen Zhang[1], Man Zhang[2],
Shibiao Xu[2,*], Weiliang Meng[1,*], and Xiaopeng Zhang[1]

***Abstract*— The complementarity between camera and LiDAR data makes fusion methods a promising approach to improve 3D semantic segmentation performance. Recent transformer-based methods have also demonstrated superiority in segmentation. However, multimodal solutions incorporating transformers are underexplored and face two key inherent difficulties: over-attention and noise from different modal data. To overcome these challenges, we propose a Deformable Multimodal Representation Fusion (DefFusion) framework consisting mainly of a Deformable Representation Fusion Transformer and Dynamic Representation Augmentation Modules. The Deformable Representation Fusion Transformer introduces the deformable mechanism in multimodal fusion, avoiding over-attention and improving efficiency by adaptively modeling a 2D key/value set for a given 3D query, thus enabling multimodal fusion with higher flexibility. To enhance the 2D representation and 3D representation, the Dynamic Representation Enhancement Module is proposed to dynamically remove noise in the input representation via Dynamic Grouped Representation Generation and Dynamic Mask Generation. Extensive experiments validate that our model achieves the best 3D semantic segmentation performance on SemanticKITTI and NuScenes benchmarks.**

I. Introduction

Semantic segmentation is vital for scene understanding and has broad applications in areas such as autonomous driving and robotics [1], [2], [3], [4]. While single modality methods for 3D semantic segmentation that using camera images [5] or LiDAR point clouds [6] as input are rapidly developing, they face unavoidable limitations due to the inherent constraints of single-modal data in complex scenes. For example, cameras can capture a wealth of color and texture details that reveal object appearance, but they are vulnerable to variations in illumination and object sizes. Moreover, depth perception from camera images can often be ambiguous, while LiDAR point clouds provide precise depth information, but the laser spots produced are too sparse to capture object details. Due to the complementary nature of cameras and LiDAR, a multimodal fusion approach that combines both is a promising method for improving the performance of 3D semantic segmentation [7], [8].

Recently, many 3D semantic segmentation methods have adopted the use of transformers to model global feature relationships [9]. Transformers are adept at modeling long-distance dependencies [10]. However, the use of transformers in multimodal segmentation methods has not been fully explored. Previous methods often carry out fusion by projecting point clouds onto image planes and then fusing corresponding image features with point features based on point-to-pixel mapping [7], [11]. Alternatively, some methods use knowledge distillation [12] to transfer useful information between different modalities [8], [13]. However, these methods do not make full use of the transformer's ability to handle multimodal data. Directly applying transformers to multi-modal fusion semantic segmentation will bring the following unavoidable limitations: 1) The challenge of over-attention brought by transformers. Fusion-based methods simultaneously processing images and point clouds consume more computing resources, which together with over-attention can bring a huge burden to the application. 2) Noise brought by different modality data during fusion. Since the camera and LiDAR are affected by the environment, deviations and errors may occur in the multimodal data.

To address the aforementioned challenges, we propose a Deformable Multimodal Representation Fusion (DefFusion) framework from the perspective of deformable representation fusion and multimodal representation enhancement. Our DefFusion framework is comprised of two main components: the Deformable Representation Fusion Transformer and the Dynamic Representation Enhancement Module. To accurately model and capture the relationships between multi-modal tokens while simultaneously preventing over-attention, we design the Deformable Representation Fusion Transformer to transfer keys and values to specific key regions within 2D representations. To achieve this, we leveraged 3D representation-guided offsets through the implementation of several sets of learnable deformable mechanisms, allowing for the generation of reference points on 2D representations with respect to 3D queries. Ultimately, this process generates corresponding learnable offsets for all reference points within the query, resulting in a flexible 2D candidate key/value set. This mechanism provides the required flexibility and efficiency needed to effectively perform multimodal fusion.

To avoid the noise of multimodal data during fusion, we provide the Dynamic Representation Enhancement Module to dynamically remove the noise in the input representation via Dynamic Grouped Representation Generation and Dynamic Mask Generation. Specifically, we cluster the channels of obtained representations into N region-aware feature maps (grouped representations). To further remove the noise of grouped representations, we learn the dynamic masks corresponding to each region via Dynamic Mask Generation.

In total, our main contributions are summarized as follows:

- We propose the Deformable Multimodal Representation

[1] State Key Laboratory of Multimodal Artifical Intelligence Systems, CASIA; [2] School of Artificial Intelligence, BUPT.
[*] Shibiao Xu and Weiliang Meng are the corresponding authors (shibiaoxu@bupt.edu.cn; weiliang.meng@ia.ac.cn).

Fusion (DefFusion) framework for 3D semantic segmentation, which demonstrates state-of-the-art performance.
- We design the Deformable Representation Fusion Transformer, which addresses over-attention with deformable representation generation, endowing multi-modal fusion with higher flexibility and efficiency.
- We provide the Dynamic Representation Enhancement Module, which effectively denoises and enhances image or point cloud representation.

II. Related Work

A. Single-Sensor 3D Semantic Segmentation

Camera-based Methods. With the development of deep learning [14], [15], [16], [17], [18], [19], many camera-based semantic segmentation methods [20] are proposed to improve FCNs [21], such as PSPNet [22] and DeeplabV3 [5] that adopt a pyramid structure. However, camera-based methods are not robust in low-light conditions compared to LiDAR-based methods.

LiDAR-based Methods. LiDAR-only 3D semantic segmentation methods typically employ the U-Net [23] architecture. There are three prevalent approaches: **i) Point.** Point methods derived from PointNet++ [24] become computationally expensive for processing large-scale LiDAR point clouds. **ii) Projection.** PolarNet [25], Deep LiDAR [26], and RangeNet++ [27] adopt this strategy. However, projecting 3D data to 2D results in information loss and performance degradation. **iii) 3D Voxel.** SparseConv [1] stores only non-empty voxels in a hash table, which improves efficiency. Recently, fusion methods [28], [6] that utilize multiple representations (points, projection, and voxels) have emerged. However, these methods only take radar point clouds as input, lacking appearance and texture information from camera images.

B. Multi-modal 3D Semantic Segmentation

Multimodal 3D semantic segmentation aims to leverage the complementary information from cameras and LiDAR. Existing methods for fusing multimodal representations include RGBAL [11], which transforms RGB images into polar grid map representations, and PMF [7], which projects point clouds onto the camera image plane for fusion. However, these methods have limitations in enhancing the point cloud due to the image format, which may hinder generalization in practical applications. Recently, new methods such as 2DPASS [8] have utilized multi-scale fusion-to-single knowledge distillation to transfer 2D semantic information into 3D networks, and Hou et al. [13] propose point-to-voxel knowledge distillation (PVD) to transfer hidden knowledge. In contrast, our approach performs both image representation augmentation and point cloud representation augmentation in the segmentation network, and achieves efficient and flexible fusion of multimodal representations using Deformable Representation Fusion Transformer.

C. Deformable mechanism

Deformable convolution [29] has been widely applied in computer vision. Deformable DETR [30] employs deformable attention to select a small number of keys for each query on top of a CNN backbone, while DPT [31] and PS-ViT [32] refine visual tokens with deformable modules. However, none of these works integrate deformable mechanisms into multimodal fusion. Our approach uses a powerful yet simple design to learn 3D-guided offsets and transfer keys and values to important regions in 2D representations for efficient fusion. Our method can also be viewed as an adaptive fusion mechanism.

III. Method

A. Framework Overview

As illustrated in Figure 1, our DefFusion comprises the Deformable Representation Fusion Transformer (DRFT), the Dynamic Representation Enhancement Module (DREM), a 2D backbone, and a 3D backbone. The DRFT utilizes deformable mechanisms to efficiently fuse multimodal representations. The DREM dynamically removes noise in 2D/3D representation to enhance feature representation. Specifically, the DREM-2D enhances 2D image representation, while the DREM-3D enhances 3D point cloud representation.

Let $X_{in} \in \mathbb{R}^{N_{camera} \times H^c \times W^c \times 3}$ represent the RGB image, and $P_{in} \in \mathbb{R}^{N_{point} \times D}$ represent the LiDAR point cloud. We feed P_{in} and the sampled X_{in} into the 2D and 3D backbones, respectively. The backbone networks extract two sets of multi-scale feature maps to obtain 2D and 3D representations, denoted by $\left\{F_l^{2D}\right\}_{l=1}^{L}$ and $\left\{F_l^{3D}\right\}_{l=1}^{L}$, where L denotes the number of stages, which is set to 4 in our framework.

To extract 2D features, we employ the HRNet-w48 [33] encoder. The last two stages of the encoder are enhanced with DREM-2D to improve the quality of the 2D representations $\left\{F_3^{2D}, F_4^{2D}\right\}$. At each stage, we use the DRFT to efficiently fuse 2D and 3D representations, which are concatenated to produce multi-scale fusion representations. Finally, we use the FCN [34] decoder to upsample the multi-scale fusion representations.

For the 3D backbone, we employ a modified SPVCNN [6] to build the 3D encoder. Similar to the 2D branch, we also apply the DREM-3D to enhance the last two stages of the 3D encoder, resulting in improved 3D representations $\left\{F_3^{3D}, F_4^{3D}\right\}$. The 3D representations of different scales $\left\{F_l^{3D}\right\}_{l=1}^{L}$ are then upsampled to the original size and concatenated with the multi-scale fusion representations. The resulting feature maps are fed into a decoder that comprises several fully connected layers. The 2D backbone and 3D backbone can be flexibly selected from various mature networks, which is verified in Section IV-C.

B. Deformable Representation Fusion Transformer

Multimodal semantic segmentation methods combined with transformers often face over-attention challenges due to input data of different modalities as query, key, and value,

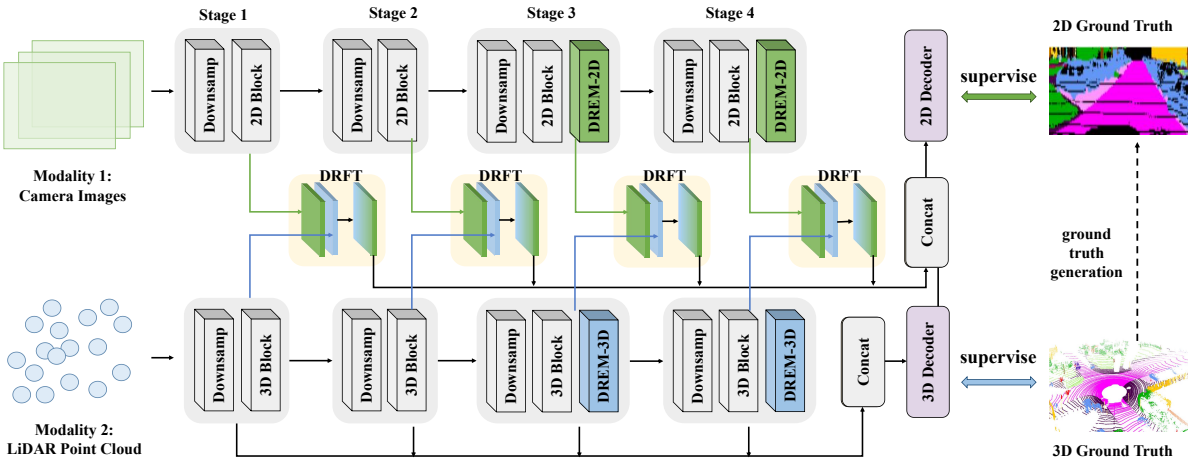

Fig. 1: Overview of our DefFusion. Images and LiDAR point clouds are respectively fed into 2D and 3D encoders to generate multi-scale features. Then, Deformable Representation Fusion Transform is applied for each scale to effectively model the relationship between multimodal tokens, thereby overcoming over-attention. DREM-2D and DREM-3D are applied in the last two stages of the 2D encoder and 3D encoder respectively to remove noise and enhance feature representation.

or directly feeding fused representations into the transformer. This exacerbates the noise that naturally arises from different modalities and can harm performance. The DCN [35] gives deformable mechanisms that using data-dependent sparse attention to solve over-attention, but previous methods have not fully utilized deformable mechanisms for multimodal representation fusion. Additionally, directly implementing deformable mechanisms in transformers poses a challenge due to the sharp rise in space complexity when learning offsets for each feature map element. In contrast, we propose a flexible approach that leverages 3D representations to guide the effective fusion of important regions in 2D representations, enabling efficient and adaptive fusion and generating semantically rich fused representations that boost segmentation performance.

Specifically, to flexibly fuse multimodal representations and effectively model the relationship between tokens, we design the Deformable Representation Fusion Transformer (Figure 2). The key idea is to use an adaptive set of deformable points to determine the important regions in the 2D representation . These points are learned by the offset network from queries $\mathbf{Q}$ obtained from 3D representations. To obtain deformed key $\hat{\mathbf{K}}$ and deformed value $\hat{\mathbf{V}}$, we perform key projection and value projection on the sampled 2D representations using the grid sample operation. Finally, we feed $\mathbf{Q}$, $\hat{\mathbf{K}}$, and $\hat{\mathbf{V}}$ into Multi-head Attention with Bias to obtain the fusion representation.

Deformable Representation Generation. To improve computational efficiency, we adopt the Point-to-Pixel Correspondence method proposed in [8] to generate paired features for both 2D and 3D multi-scale representations. These paired features are then used as the input representations $\left\{X_l^{2D}\right\}_{l=1}^{L}$ and $\left\{X_l^{3D}\right\}_{l=1}^{L}$ for the Deformable Representation Fusion Transformer in the i-th stage of the encoder.

For generating a deformable representation from a 2D input representation $\mathbf{X}_l^{2D} \in \mathbb{R}^{N_l \times C_l}$, we first initialize a set of reference points $\mathbf{p}$ and downsample the grid size by a factor r from the input representation size, resulting in $N_l^s = N_l/r$ points. The reference points are assigned values of linearly spaced coordinates ranging from 0 to $N_l^s - 1$. To obtain the offset of each reference point, we linearly project the 3D representation $X_l^{3D} \in \mathbb{R}^{N_l \times C_l}$ to the query Q. We then apply a lightweight offset network $\theta_{offset}(\cdot)$ to the query to obtain the offsets $\Delta\mathbf{p}$. The deformable representation generation process can be expressed as:

$$\mathbf{Q} = \varphi_q(\mathbf{X_l^{3D}}), \hat{\mathbf{K}} = \varphi_k(\hat{\mathbf{X}}^{\mathbf{2D}}), \hat{\mathbf{V}} = \varphi_v(\hat{\mathbf{X}}^{\mathbf{2D}}) \quad (1)$$

$$\Delta\mathbf{p} = \theta_{offset}(\mathbf{Q}), \quad \hat{\mathbf{X}}^{\mathbf{2D}} = s(\mathbf{X}_l^{2D}; \mathbf{p} + \Delta\mathbf{p}) \quad (2)$$

where $\hat{\mathbf{K}}$ and $\hat{\mathbf{V}}$ denote the deformable key embedding and deformable value embedding guided by the 3D representation, respectively. $s(\cdot)$ represents the grid sampling function, which is set to a bilinear interpolation. To learn reasonable offsets, we implement the offset network $\theta_{offset}(\cdot)$ as two linear layers with nonlinear activation, where we employ the ReLU activation function.

Multi-head Attention with Bias. To efficiently compute self-attention, we employ multi-head attention on $\mathbf{Q}$, $\hat{\mathbf{K}}$, and $\hat{\mathbf{V}}$. We also adopt a relative position bias $\mathbf{B} \in \mathbb{R}^{N_l \times N_l^s}$ when computing attention. The relative positional bias $\mathbf{B}$ is equivalent to the position embedding used in [10] and supplements spatial information by encoding the relative position between each pair of query and key. As a result, the output of one head is expressed as:

$$\mathbf{F}_m^{attn} = Softmax(\omega_m)\mathbf{V}_m \quad (3)$$

$$\omega_m = \gamma(\mathbf{Q}_m, \hat{\mathbf{K}}_m)/\sqrt{d} + \mathbf{B} \quad (4)$$

where the attention weight ω_m is scalars computed from the scaled dot-product $\gamma(\cdot)$. d is the dimension of each head. The attention features $\mathbf{f}_m^{attn}$ obtained by multi-head attention are

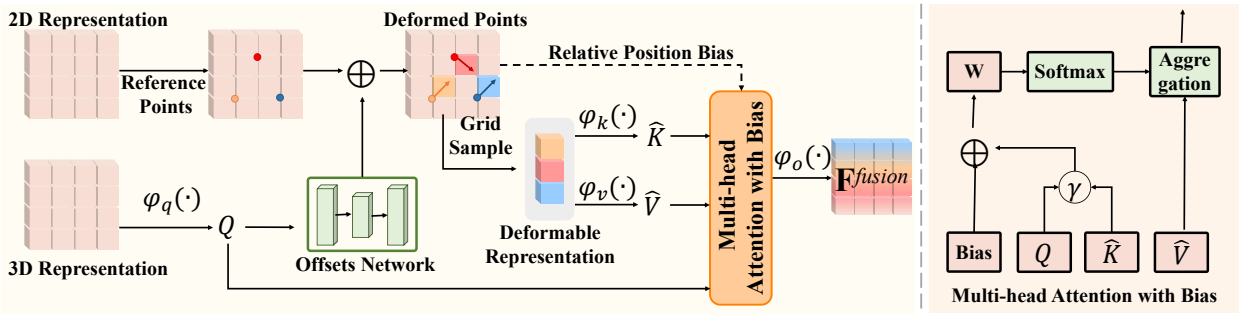

Fig. 2: The illustration of our Deformable Representation Fusion Transformer. A set of reference points are uniformly placed on the 2D representation, and the offsets are learned from the 3D query through an offset network. Then, the deformable keys and values are projected from the deformable representation based on the deformable points. The relative positional deviation is also calculated by the deformable points, which enhances the multi-head attention.

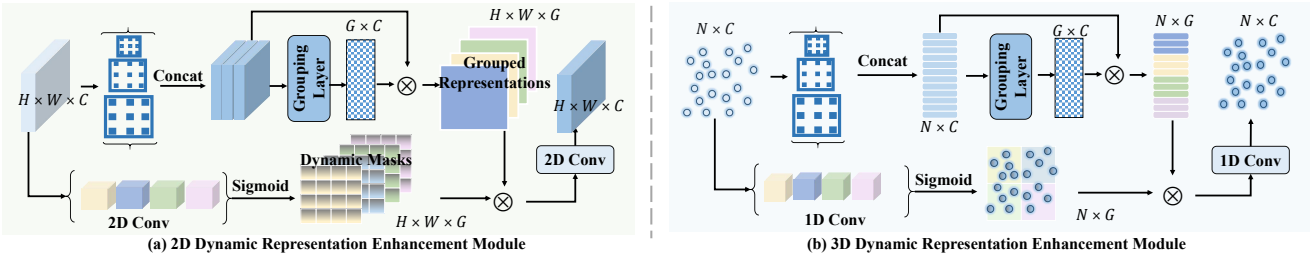

Fig. 3: The illustration of our Dynamic Representation Enhancement Module. (a) DREM-2D: For the input representation, DREM-2D generates grouped representations and dynamic masks, and multiplies them element by element to obtain an enhanced representation. (b) DREM-3D: The operation of DREM-3D is parallel to DREM-2D.

linearly projected via φ_o to get the final output $\mathbf{f}^{fusion}$:

$$\mathbf{F}^{fusion} = \varphi_o(Concat(\mathbf{F}_1^{attn}, \ldots, \mathbf{F}_m^{attn})) \tag{5}$$

In practice, our multi-head attention adopts the window-based attention in [10], [36] for computational efficiency.

C. Dynamic Representation Enhancement Module

To enhance the original representation of the 2D backbone and 3D backbone and dynamically remove the noise in the original representation, we provide the Dynamic Representation Enhancement Module as shown in Figure 3.

1) 2D Dynamic Representation Enhancement Module: **Dynamic Grouped Representation Generation.** We augment the original 2D representation with a set of dilated convolutions, denoted as $F^{2D} \in \mathbb{R}^{H \times W \times C}$. More specifically, we apply dilated convolutions with dilation rates of 3, 6, and 12, respectively. The resulting outputs are concatenated together to obtain F_e^{2D}. However, the number of channels in F_e^{2D} is typically quite large, which may lead to excessive computational costs. For computational efficiency, we follow [37] to aggregate channels of the representation F_e^{2D} using K-means, and employ the clustering results as the representation of key regions. Specifically, we cluster the channels of F_e^{2D} into G groups, representing G key regions. The clustering result is expressed as a matrix $cl \in \mathbb{R}^{G \times C}$, where $cl_{i,j}$ ranges from 0 to 1, indicating the probability that the j-th channel belongs to the i-th group (i.e. the i-th region). We employ two fully connected layers and a sigmoid activation function to approximate the clustering process, called the grouping layer (GL). We use F_e^{2D} as the input of GL to obtain the clustering result cl:

$$cl = GL(F_e^{2D}) = Sigmoid(FC(F_e^{2D})) \tag{6}$$

Then We utilize cl to calculate the dynamic grouped representation $F_g^{2D} \in \mathbb{R}^{H \times W \times G}$ of each key region. For the i-th region, the calculation formula of its corresponding representation $F_{gi}^{2D} \in \mathbb{R}^{H \times W}$ is:

$$F_{gi}^{2D} = \frac{1}{C} \sum_{c=1}^{C} cl_{ic} \times F_{ec}^{2D} \tag{7}$$

where cl_{ic} is an estimated value indicating whether the c-th channel belongs to the i-th region, and F_{ec}^{2D} is the feature of the c-th channel of F_e^{2D}. We obtain the dynamic grouped representation F_g^{2D} by concatenating features from all regions.

Dynamic mask generation. Representative representations after grouping inevitably contain noise. To further enhance the feature representation, we apply a denoising process to the F_g^{2D} by generating dynamic masks. Specifically, for the i-th group (i.e. the i-th region), we use a 2D convolution and a sigmoid activation function to learn the corresponding mask M_i, and make m_i range from 0-1. The mask M_i represents the importance of each pixel in the space of the F_{gi}^{2D}.

$$M_i = Sigmoid(Conv(F^{2D})) \tag{8}$$

TABLE I: Performance comparison on the SemanticKITTI test benchmark. The best results are marked with **bold**.

Methods	**mIoU**	road	sidewalk	parking	other-ground	building	car	truck	bicycle	motorcycle	other-vehicle	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traffic sign
SqueezeSegV2$_{ICRA'19}$ [38]	39.7	88.6	67.6	45.8	17.7	73.7	81.8	13.4	18.5	17.9	14.0	71.8	35.8	60.2	20.1	25.1	3.9	41.1	20.2	26.3
RandLA-Net$_{CVPR'20}$ [39]	55.9	90.5	74.0	61.8	24.5	89.7	94.2	43.9	29.8	32.2	39.1	83.8	63.6	68.6	48.4	47.4	9.4	60.4	51.0	50.7
PolarNet$_{CVPR'20}$ [25]	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
JS3C-Net$_{AAAI'21}$ [40]	66.0	88.9	72.1	61.9	31.9	92.5	95.8	54.3	59.3	52.9	46.0	84.5	69.8	67.9	69.5	65.4	39.9	70.8	60.7	68.7
Cylinder3D$_{ArXiv'20}$ [41]	68.9	92.2	77.0	65.0	32.3	90.7	97.1	50.8	67.6	63.8	58.5	85.6	72.5	69.8	73.7	69.2	48.0	66.5	62.4	66.2
RPVNet$_{ICCV'21}$ [28]	70.3	93.4	**80.7**	**70.3**	33.3	93.5	97.6	44.2	**68.4**	68.7	61.1	86.5	75.1	71.7	75.9	74.4	43.4	72.1	64.8	61.4
(AF)2-S3Net$_{CVPR'21}$ [42]	70.8	92.0	76.2	66.8	**45.8**	92.5	94.3	40.2	63.0	**81.4**	40.0	78.6	68.0	63.1	76.4	81.7	77.7	69.6	64.0	73.3
PVKD$_{CVPR'22}$ [13]	71.2	91.8	77.5	70.9	41.0	92.4	97.0	53.5	67.9	69.3	60.2	86.5	73.8	71.9	75.1	73.5	50.5	69.4	64.9	65.8
2DPASS$_{ECCV'22}$ [8]	72.9	89.7	74.7	67.4	40.0	93.5	97.0	61.1	63.6	63.4	61.5	86.2	73.9	71.0	77.9	81.3	74.1	72.9	65.0	70.4
Baseline	68.0	90.1	73.9	62.7	34.6	91.4	95.2	55.3	52.5	56.8	53.1	85.9	72.6	71.4	76.3	79.5	38.7	68.8	64.5	69.2
DefFusion (Ours)	**74.6**	**93.5**	76.8	68.3	42.7	**94.2**	**97.8**	**62.4**	65.0	67.1	**62.9**	**87.3**	**75.4**	**72.2**	**78.6**	**82.1**	**77.8**	**73.5**	**67.4**	**71.9**

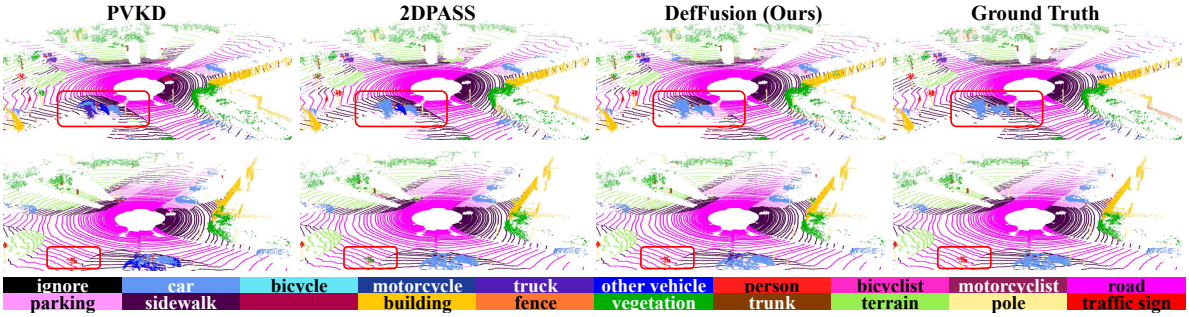

Fig. 4: Visual comparison of various methods on the SemanticKITTI validation set. It can be observed that our method performs the best, especially on challenging objects such as persons, cars, and other vehicles.

Finally, we element-wise multiply M with F_g^{2D} to obtain a denoised representative representation.

2) 3D Dynamic Representation Enhancement Module: We can easily extend the DREM-2D to the DREM-3D by replacing 2D convolution and 2D dilated convolution with 1D convolution and 1D dilated convolution, respectively. DREM-3D, like DREM-2D, can be divided into Dynamic Grouped Representation Generation and Dynamic mask generation. Specifically, to calculate the grouped representation F_{gi}^{3D}, we use 1D dilated convolution and a grouping layer (GL) to obtain clustering results. To denoise F_{gi}^{3D}, we learn a dynamic mask M_i' for the i-th region via a 1D convolution and sigmoid activation function.

IV. Experiment

A. Experimental Setup

Dataset. We use SemanticKITTI Dataset and NuScenes Dataset. The SemanticKITTI Dataset provides dense semantic annotations for scans of sequence 00-10 in the KITTI dataset [44]. The NuScenes dataset comprises 1000 scenes that encompass diverse traffic and weather conditions.

Evaluation Metric We evaluate methods using Intersection-over-Union (mIoU), which is defined as $\frac{TP}{TP+FP+FN}$, where TP, FP, FN are the number of true positive, false positive, and false negative predictions. In addition, we also used overall accuracy (Acc) for evaluation.

Settings. We adopt the semantic segmentation loss functions of cross-entropy and Lovasz loss following [8] for training our DefFusion end-to-end using the SGD optimizer. We set the number of groups G to 4 for both DREM-2D and DREM-3D. To augment the 2D input, we apply the standard techniques of horizontal flipping and color jitter. Similarly, for the 3D input, we utilize global rotation at random angles and global scaling with a random scale factor.

B. Comparison with State-of-the-art Methods

SemanticKITTI Results. We give a comprehensive evaluation of our method and other approaches on the SemanticKITTI dataset in Table I. Our DefFusion method achieves the highest mIoU compared to all other state-of-the-art methods, including single sensor-based segmentation methods and multimodal fusion-based segmentation methods. Our method significantly improves the mIoU by 9.7% compared to the baseline, which validates the effectiveness of our DRFT and DREM. In addition, when DefFusion uses the same 2D and 3D backbone as 2DPASS, it achieves a mIoU of 74.0% on SemanticKITTI.

NuScenes Results. Table II reports the semantic segmentation results of our DefFusion and other state-of-the-art methods on the NuScenes dataset. Our method not only outperforms all single sensor-based methods but also surpasses other multimodal fusion-based methods. our DefFusion achieves the best mIoU, improving it by 5.1% compared to the baseline, and achieving state-of-the-art performance with 81.8% mIoU.

Comparison of Qualitative Results. In Figure 4, we give the qualitative results of our method and other multimodal fusion-based methods on SemanticKITTI validation set.

C. Ablation Study

Effectiveness of Various Components in DefFusion. We first removed key components in DefFusion to verify

TABLE II: Performance comparison on the NuScenes test benchmark. L means use LiDAR only. $L + C$ stands for using LiDAR and cameras.

Methods	Input	**mIoU**	barrier	bicycle	bus	car	construction	motorcycle	pedestrian	traffic cone	trailer	truck	driveable	other flat	sidewalk	terrain	manmade	vegetation
PolarNet$_{CVPR'20}$ [25]	L	69.4	72.2	16.8	77.0	86.5	51.1	69.7	64.8	54.1	69.7	63.5	96.6	67.1	77.7	72.1	87.1	84.5
JS3C-Net$_{AAAI'21}$ [40]	L	73.6	80.1	26.2	87.8	84.5	55.2	72.6	71.3	66.3	76.8	71.2	96.8	64.5	76.9	74.1	87.5	86.1
PMF$_{ICCV'21}$ [7]	L+C	77.0	82.0	40.0	81.0	88.0	64.0	79.0	80.0	**76.0**	81.0	67.0	97.0	68.0	78.0	74.0	90.0	88.0
2D3DNet$_{3DV'21}$ [43]	L+C	80.0	83.0	**59.4**	88.0	85.1	63.7	84.4	**82.0**	**76.0**	84.8	71.9	96.9	67.4	79.8	76.0	**92.1**	89.2
2DPASS$_{ECCV'22}$ [8]	L+C	80.8	81.7	55.3	92.0	91.8	73.3	86.5	78.5	72.5	84.7	75.5	97.6	69.1	79.9	75.5	90.2	88.0
Baseline	L+C	77.8	80.5	38.4	91.7	90.8	66.3	79.2	70.9	71.3	83.0	75.6	96.4	69.3	79.1	75.5	89.7	87.2
DefFusion (Ours)	L+C	**81.8**	**83.2**	57.9	**92.1**	**92.2**	74.6	**87.4**	80.7	73.3	**85.3**	**76.5**	97.0	69.6	**80.9**	**76.8**	91.6	**89.3**

TABLE III: Ablation studies of our DefFusion on SemanticKITTI validation set.

Methods	DRFT	DREM-2D	DREM-3D	SemanticKITTI
Baseline				65.8
DefFusion	✔			68.3
	✔	✔		69.4
	✔	✔	✔	**70.8**

TABLE IV: Ablation studies of our DRFT on SemanticKITTI validation set. $Def.Gen.$ denotes using Deformable Representation Generation to generate deformable keys and deformable values. $Pos.Bias$ denotes using Attention with Position Bias. S represents multi-head attention. V represents vector attention in [9], W represents window-based attention in [10], [36].

Def. Gen.	Pos. Bias	Attn.	SemanticKITTI
		W	66.4
✔		W	69.0
✔	✔	S	69.2
✔	✔	V	69.3
✔	✔	W	**70.8**

the validity of these designs. Table III reports the ablation studies of DefFusion on the SemanticKITTI validation set. Compared with the baseline, our DRFT further significantly improves the mIoU by 3.8%, demonstrating the importance of efficiently fusing multimodal representations with deformable mechanisms. Moreover, applying DREM-2D and DREM-3D enables the framework to achieve 69.4% and 70.8% mIoU, respectively.

Exploring of Deformable Representation Fusion Transformer. We show the performance of deleting key components in the DRFT in Table IV. Using Deformable Representation Generation ($Def.Gen.$) or using Attention with Relative Position Bias ($Pos.Bias$) in feature sampling provides a significant improvement. In addition, we explore various types of attention heads [9], [10], [36]. Among these, we adopt the best-performing window-based attention as our attention head.

Robustness Against Camera Malfunction. We evaluate the robustness of our DefFusion by examining its performance under adverse conditions, such as camera failures, as shown in Table V. Remarkably, our DefFusion technique still delivers impressive results even when all cameras are removed, surpassing the LiDAR-only baseline.

TABLE V: Robustness analysis on DefFusion on the NuScenes validation set by removing some cameras as malfunctions. $Camera$: the number of available cameras.

$Camera$	LiDAR-only	0	1	2	3	4	5	6
mIoU (%)	71.7	74.6	75.5	76.1	77.0	77.8	79.0	80.4

TABLE VI: Generality analysis by varying the backbone settings on the SemanticKITTI validation set.

Backbone	DRFT	DREM	SemanticKITTI	$gain$
MinkUNet [6]			63.8	
	✔		67.3	+3.5
	✔	✔	69.3	+5.5
SegFormer-B5 [45]			64.7	
	✔		67.7	+3.0
	✔	✔	70.0	+5.3

Generality Analysis. Table VI illustrates that our DefFusion method can enhance segmentation performance through a "backbone-independent" training approach. We conduct experiments on two additional baseline models, the MinkUNet [6] replacing the 3D backbone (SPVCNN), and the SegFormer-B5 [45] replacing the 2D backbone (HRNet-w48). All other settings remained constant except for the backbone-related components. As shown in Table VI, our DRFT and DREM achieved significant improvements in segmentation performance on different backbones, indicating the efficacy and versatility of our DefFusion technique.

V. Conclusion

In this paper, we propose a Deformable Multimodal Representation Fusion (DefFusion) framework for 3D semantic segmentation to address the challenges of over-attention brought by transformers and noise brought by different modal data. The DefFusion employ our designed Deformable Representation Fusion Transformer to adaptively model the relationship between multi-modal tokens and avoid over-attention. Furthermore, we provide Dynamic Representation Enhancement Module to dynamically removes noise in the input representation via Dynamic Grouped Representation Generation and Dynamic Mask Generation, which is suitable for both 2D and 3D representation enhancement. Eventually, our method achieves state-of-the-art on two large-scale recognized benchmarks, namely SemanticKITTI and NuSchenes.

VI. Acknowledgements

This work was supported by Beijing Natural Science Foundation (No. JQ23014 and No. L231013), and by the National Natural Science Foundation of China (Nos. U21A20515, 62271074 and 62376271).

References

[1] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9224–9232.

[2] R. Xu, C. Wang, J. Sun, S. Xu, W. Meng, and X. Zhang, "Self correspondence distillation for end-to-end weakly-supervised semantic segmentation," *arXiv preprint arXiv:2302.13765*, 2023.

[3] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation," *IEEE Transactions on Image Processing*, 2023.

[4] R. Xu, C. Wang, S. Xu, W. Meng, and X. Zhang, "Dc-net: Dual context network for 2d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 503–513.

[5] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[6] H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching efficient 3d architectures with sparse point-voxel convolution," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII*. Springer, 2020, pp. 685–702.

[7] Z. Zhuang, R. Li, K. Jia, Q. Wang, Y. Li, and M. Tan, "Perception-aware multi-sensor fusion for 3d lidar semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 280–16 290.

[8] X. Yan, J. Gao, C. Zheng, C. Zheng, R. Zhang, S. Cui, and Z. Li, "2dpass: 2d priors assisted semantic segmentation on lidar point clouds," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*. Springer, 2022, pp. 677–695.

[9] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 259–16 268.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[11] K. El Madawi, H. Rashed, A. El Sallab, O. Nasr, H. Kamel, and S. Yogamani, "Rgb and lidar fusion based 3d semantic segmentation for autonomous driving," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 7–12.

[12] D. Zhang, H. Li, W. Cong, R. Xu, J. Dong, and X. Chen, "Task relation distillation and prototypical pseudo label for incremental named entity recognition," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 3319–3329.

[13] Y. Hou, X. Zhu, Y. Ma, C. C. Loy, and Y. Li, "Point-to-voxel knowledge distillation for lidar semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8479–8488.

[14] M. Zhao, W. Yan, R. Xu, D. Zhi, R. Jiang, T. Jiang, V. D. Calhoun, and J. Sui, "An attention-based hybrid deep learning framework integrating temporal coherence and dynamics for discriminating schizophrenia," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 118–121.

[15] D. Zhang, Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu, "Mm-llms: Recent advances in multimodal large language models," *arXiv preprint arXiv:2401.13601*, 2024.

[16] R. Xu, C. Wang, S. Xu, W. Meng, Y. Zhang, B. Fan, and X. Zhang, "Domainfeat: Learning local features with domain adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[17] R. Xu, C. Wang, S. Xu, W. Meng, and X. Zhang, "Dual-stream representation fusion learning for accurate medical image segmentation," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106402, 2023.

[18] Y. Zhang, Y. Liu, D. Miao, Q. Zhang, Y. Shi, and L. Hu, "Mg-vit: A multi-granularity method for compact and efficient vision transformers," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[19] L.-H. Zhao, J.-L. Wang, and Y. Zhang, "Lag output synchronization for multiple output coupled complex networks with positive semidefinite or positive definite output matrix," *Journal of the Franklin Institute*, vol. 357, no. 1, pp. 414–436, 2020.

[20] R. Xu, C. Wang, S. Xu, W. Meng, and X. Zhang, "Wave-like class activation map with representation fusion for weakly-supervised semantic segmentation," *IEEE Transactions on Multimedia*, vol. 26, pp. 581–592, 2024.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

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

[25] Y. Zhang, Z. Zhou, P. David, X. Yue, Z. Xi, B. Gong, and H. Foroosh, "Polarnet: An improved grid representation for online lidar point clouds semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9601–9610.

[26] F. J. Lawin, M. Danelljan, P. Tosteberg, G. Bhat, F. S. Khan, and M. Felsberg, "Deep projective 3d semantic segmentation," in *Computer Analysis of Images and Patterns: 17th International Conference, CAIP 2017, Ystad, Sweden, August 22-24, 2017, Proceedings, Part I 17*. Springer, 2017, pp. 95–107.

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

[28] J. Xu, R. Zhang, J. Dou, Y. Zhu, J. Sun, and S. Pu, "Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 024–16 033.

[29] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9308–9316.

[30] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[31] Z. Chen, Y. Zhu, C. Zhao, G. Hu, W. Zeng, J. Wang, and M. Tang, "Dpt: Deformable patch-based transformer for visual recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2899–2907.

[32] X. Yue, S. Sun, Z. Kuang, M. Wei, P. H. Torr, W. Zhang, and D. Lin, "Vision transformer with progressive sampling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 387–396.

[33] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[34] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[35] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[36] X. Lai, J. Liu, L. Jiang, L. Wang, H. Zhao, S. Liu, X. Qi, and J. Jia, "Stratified transformer for 3d point cloud segmentation," in

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8500–8509.

[37] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5209–5217.

[38] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4376–4382.

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

[40] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3101–3109.

[41] H. Zhou, X. Zhu, X. Song, Y. Ma, Z. Wang, H. Li, and D. Lin, "Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation," *arXiv preprint arXiv:2008.01550*, 2020.

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

[43] K. Genova, X. Yin, A. Kundu, C. Pantofaru, F. Cole, A. Sud, B. Brewington, B. Shucker, and T. Funkhouser, "Learning 3d semantic segmentation with only 2d image supervision," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 361–372.

[44] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[45] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.