

# CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers

Jiaming Zhang\*, Huayao Liu\*, Kailun Yang\*<sup>†</sup>, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen

arXiv:2203.04838v5 [cs.CV] 24 Nov 2023

**Abstract**—Scene understanding based on image segmentation is a crucial component of autonomous vehicles. Pixel-wise semantic segmentation of RGB images can be advanced by exploiting complementary features from the supplementary modality ( $X$ -modality). However, covering a wide variety of sensors with a modality-agnostic model remains an unresolved problem due to variations in sensor characteristics among different modalities. Unlike previous modality-specific methods, in this work, we propose a unified fusion framework, *CMX*, for RGB-X semantic segmentation. To generalize well across different modalities, that often include supplements as well as uncertainties, a unified cross-modal interaction is crucial for modality fusion. Specifically, we design a Cross-Modal Feature Rectification Module (*CM-FRM*) to calibrate bi-modal features by leveraging the features from one modality to rectify the features of the other modality. With rectified feature pairs, we deploy a Feature Fusion Module (*FFM*) to perform sufficient exchange of long-range contexts before mixing. To verify *CMX*, for the first time, we unify five modalities complementary to RGB, *i.e.*, depth, thermal, polarization, event, and LiDAR. Extensive experiments show that *CMX* generalizes well to diverse multi-modal fusion, achieving state-of-the-art performances on five RGB-Depth benchmarks, as well as RGB-Thermal, RGB-Polarization, and RGB-LiDAR datasets. Besides, to investigate the generalizability to dense-sparse data fusion, we establish an RGB-Event semantic segmentation benchmark based on the EventScape dataset, on which *CMX* sets the new state-of-the-art. The source code of *CMX* is publicly available at [https://github.com/huaaliu/RGBX\\_Semantic\\_Segmentation](https://github.com/huaaliu/RGBX_Semantic_Segmentation).

**Index Terms**—Semantic Segmentation, Scene Parsing, Cross-Modal Fusion, Vision Transformers, Scene Understanding.

## I. INTRODUCTION

**S**CENE understanding is a fundamental component in Autonomous Vehicles (AVs) since it can provide comprehensive information to support the Advanced Driver-Assistance System (ADAS) to make correct decisions when interacting with the driving surrounding [1]. As exteroceptive sensors, cameras are adopted in AVs for perceiving the surroundings [2]. Image semantic segmentation – a fundamental task in

This work was supported in part by the Federal Ministry of Labor and Social Affairs (BMAS) through the AccessibleMaps project under Grant 01KM15112, in part by the “KIT Future Fields” project, in part by the MWK through the Cooperative Graduate School Accessibility through AI-based Assistive Technology (KATE) under Grant BW6-03, and in part by the BMBF through a fellowship within the IFI program of the German Academic Exchange Service (DAAD), in part by the HoreKA@KIT supercomputer partition, and in part by Hangzhou SurImage Technology Company Ltd.

J. Zhang, R. Liu, and R. Stiefelhagen are with Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany.

K. Yang is with Hunan University, Changsha 410082, China.

H. Liu is with NIO, Shanghai 201804, China.

X. Hu is with ByteDance Inc., Hangzhou 310000, China.

\*indicates equal contribution.

<sup>†</sup>corresponding author. (E-Mail: kailun.yang@hnu.edu.cn.)

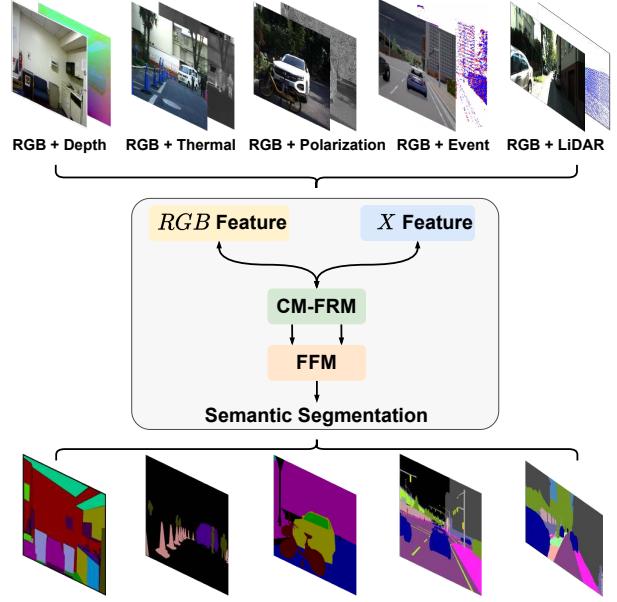


Fig. 1: RGB-X semantic segmentation unifies diverse sensing modality combinations: RGB-Depth, -Thermal, -Polarization, -Event, and -LiDAR segmentation. *CMX* is established with Cross-Modal Feature Rectification Module (*CM-FRM*) to calibrate the features of RGB- and  $X$ -modality and Feature Fusion Module (*FFM*) to perform the exchange of long-range context and combine features for RGB-X semantic segmentation.

computer vision – is an ideal perception solution to transform an image input into its underlying semantically meaningful regions, providing pixel-wise dense scene understanding for Intelligent Transportation Systems (ITS) [3], [4]. Image semantic segmentation has made significant progress on accuracy [5], [6], [7]. Yet, current models may struggle to extract high-quality features in certain circumstances, *e.g.*, when two objects have similar colors or textures, leading to difficulty in distinguishing them through pure RGB images [8].

Thanks to the development of sensor technologies, there is a growing variety of modular sensors which are highly applicable for ITS applications. Different types of sensors can supply RGB images with rich complementary information (see Fig. 1). For example, *depth* measurement can help identify the boundaries of objects and offer geometric information of dense scene elements [8], [9]. *Thermal* images facilitate to discern different objects through their specific infrared imaging [10], [11]. Besides, *polarimetric*- and *event* information are advantageous for perception in specular- and dynamic real-world

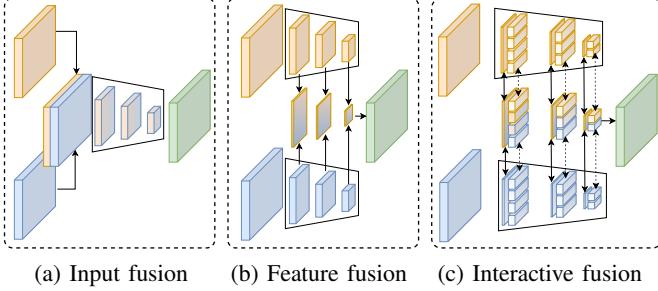


Fig. 2: Comparison of different fusion methods. (a) Input fusion merges inputs with modality-specific operations [15], [16]. (b) Feature fusion applies channel attention to fuse features in a unidirectional manner [8], [9]. (c) Our interactive fusion incorporates bidirectional cross-modal feature rectification, and sequence-to-sequence cross-attention, yielding comprehensive cross-modal interactions.

scenes [12], [13]. LiDAR data can provide spatial information in driving scenarios [14]. Thereby, a research question arises: *How to construct a unified model to incorporate the fusion of RGB with various modalities, i.e., RGB-X semantic segmentation as illustrated in Fig. 1?*

Existing multi-modal semantic segmentation methods can be divided into two categories: (1) The first category [15], [16] employs a single network to extract features from RGB and another modality, which are fused in the input stage (see Fig. 2a). (2) The second type of approaches [9], [11], [17] deploys two backbones to perform feature extraction from RGB- and another modality separately then fuses the extracted two features into one feature for semantic prediction (see Fig. 2b). However, both types are usually well-tailored for a single specific modality pair (*e.g.*, RGB-D or RGB-T), yet hard to be extended to operate with other modality combinations. For example, regarding our observation in Fig. 3, ACNet [8] and SA-Gate [9], designed for RGB-D data, perform less satisfactorily in RGB-T tasks. To flexibly cover various sensor combinations for ITS applications, a unified *RGB-X semantic segmentation*, is desirable and advantageous. Its benefits are two-fold: (1) It can save research and engineering efforts, with no need to adapt architectures for a specific modality combination scenario. (2) It makes it possible that a system equipped with multi-modal sensors can readily leverage new sensors when they become available [18], [19], which is conducive to robust scene perception. For this purpose, in this work, we spend efforts to construct a modality-agnostic framework for unified RGB-X semantic segmentation.

Recently, vision transformers [20], [21], [22], [23] handle inputs as sequences and are able to acquire long-range correlations, offering the possibility for a unified framework for diverse multi-modal tasks. Compared to existing multi-modal fusion modules [8], [12], [17] based on Convolutional Neural Networks (CNNs), it remains unclear whether potential improvements on RGB-X semantic segmentation can be materialized via vision transformers. Crucially, while some previous works [8], [9] use a simple global multi-modal interaction strategy, it does not generalize well across different

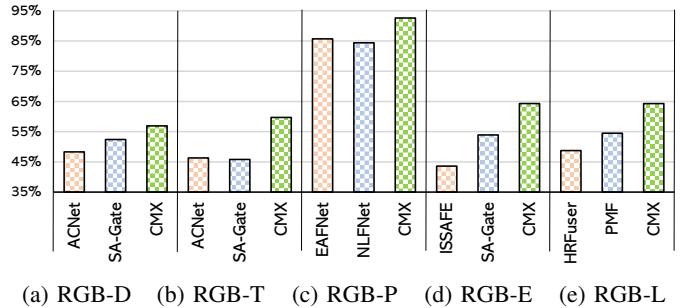


Fig. 3: Performance comparison on different RGB-X semantic segmentation benchmarks. SA-Gate [9] designed for RGB-D data (*e.g.*, on NYU Depth V2 dataset [24]), is less effective on RGB-T or RGB-E tasks. Our modality-agnostic CMX, for the first time, outperforms modality-specific methods on five segmentation tasks.

sensing data combinations [11]. We hypothesize that for RGB-X semantic segmentation with various supplements and uncertainties, comprehensive cross-modal interactions should be provided, to fully exploit the potential of cross-modal complementary features.

To tackle the aforementioned challenges, we propose *CMX*, a universal cross-modal fusion framework for RGB-X semantic segmentation in an interactive fusion manner (Fig. 2c). Specifically, CMX is built as a two-stream architecture, *i.e.*, RGB- and X-modal stream. Two specific modules are designed for feature interaction and feature fusion in between. (1) *Cross-Modal Feature Rectification Module (CM-FRM)*, calibrates the bi-modal features by leveraging their spatial- and channel-wise correlations, which enables both streams to focus more on the complementary informative cues from each other, as well as mitigates the effects of uncertainties and noisy measurements from different modalities. Such a feature rectification tackles varying noises and uncertainties in diverse modalities. It enables better multi-modal feature extraction and interaction. (2) *Feature Fusion Module (FFM)*, is constructed in two stages and it performs sufficient information exchange before merging features. Motivated by the large receptive fields obtained via self-attention [20], a cross-attention mechanism is devised in the first stage of FFM for realizing cross-modal global reasoning. In its second stage, mixed channel embedding is applied to produce enhanced output features. Thereby, our introduced comprehensive interactions lie in multiple levels (see Fig. 2c). It includes channel- and spatial-wise rectification from the feature map perspective, as well as cross-attention from the sequence-to-sequence perspective, which are critical for generalization across modality combinations.

To verify our unification proposal, we consider and assess CMX on 5 different multi-modal semantic segmentation tasks, including RGB-Depth, -Thermal, -Polarization, -Event, and -LiDAR semantic segmentation. A total of 9 datasets are involved. In particular, CMX attains top mIoU of 56.9% on NYU Depth V2 (RGB-D) [24], 59.7% on MFNet (RGB-T) [10], 92.6% on ZJU-RGB-P (RGB-P) [12], and 64.3% on KITTI-360 (RGB-L) [25] datasets. Our universal approach

CMX clearly outperforms specialized architectures (Fig. 3). Furthermore, to address the lack of RGB-Event parsing benchmark in the community, we establish an RGB-Event semantic segmentation benchmark based on the EventScape dataset [26], where our CMX sets the new state-of-the-art among >10 benchmarked models. Besides, our experiments demonstrate that the CMX framework is effective for both CNN- and Transformer-based architectures. Moreover, our investigation on representations of polarization- and event-based data indicates the path to follow and the sweet spot for reaching robust multi-modal semantic segmentation, trumping original representation methods [12], [26].

At a glance, we deliver the following contributions:

- For the first time, we explore ***RGB-X semantic segmentation*** in five types of multi-modal sensing data combinations, including **RGB-Depth**, **RGB-Thermal**, **RGB-Polarization**, **RGB-Event**, and **RGB-LiDAR**.
- We rethink multi-modality fusion from a generalization perspective and prove that comprehensive cross-modal interaction is crucial for the unification of fusion across diverse modalities.
- We propose an **RGB-X semantic segmentation framework CMX** with ***cross-modal feature rectification*** and ***feature fusion*** modules, intertwining ***cross-attention*** and ***mixed channel embedding*** for enhanced global reasoning.
- We investigate different representations of polarimetric- and event data and indicate the optimal path to follow for reaching robust multi-modal semantic segmentation.
- An RGB-Event semantic segmentation benchmark is established to assess dense-sparse data fusion, and is incorporated into the RGB-X semantic segmentation.

## II. RELATED WORK

### A. Transformer-driven Semantic Segmentation

For dense semantic segmentation, pyramid-, strip-, and atrous spatial pyramid pooling are designed to harvest multi-scale feature representations [5], [6]. Besides, cross-image pixel contrast learning [27] is applied to address intra-class compactness and inter-class dispersion, while nonparametric nearest prototype retrieving [28] is proposed to achieve semantic segmentation in a prototype view. Inspired by the non-local block [29], self-attention in transformers [20] has been used to establish long-range dependencies by DANet [7] and CCNet [30]. Recently, SETR [31] and Segmenter [32] directly adopt vision transformers [21], [22] as the backbone, which captures global context from very early layers. SegFormer [33] and Swin [23] create hierarchical structures to make use of multi-resolution features. Following this trend, various architectures of dense prediction transformers [34], [35] and semantic segmentation transformers [36], [37] emerge in the field. While these approaches have achieved high performance, most of them focus on using RGB images and suffer when RGB images cannot provide sufficient information in real-world scenes, *e.g.*, under low-illumination conditions or in high-dynamic areas. In this work, we tackle multi-modal semantic segmentation to take advantage of complementary information from other modalities such as depth, thermal, polarization, event, and LiDAR data for boosting RGB segmentation.

### B. Multi-modal Semantic Segmentation

While previous works reach high performance on standard RGB-based semantic segmentation benchmarks, in challenging real-world conditions, it is desirable to involve multi-modality sensing for a reliable and comprehensive scene understanding. **RGB-Depth** [38], [39] and **RGB-Thermal** [40], [41], [42] semantic segmentation are broadly investigated. Polarimetric optical cues [43] and event-driven priors [44] are often intertwined for robust perception under adverse conditions. In automated driving, LiDAR data [14] is incorporated for enhanced semantic road scene understanding. However, most of these works only address a single modality combination. In this work, we explore a unified approach, which can generalize well to diverse multi-modal combinations.

For multi-modal semantic segmentation, there are two dominant strategies. The first mainstream paradigm models cross-modal complementary information into layer- or operator designs [15], [16], [45], [46], [47]. While these works verify that multi-modal features can be learned within a shared network, they are carefully designed for a single modality, *e.g.*, RGB-D semantic segmentation, which is hard to be applied to other modalities. Moreover, there are multi-task frameworks [48], [49] that facilitate inter-task feature propagation for RGB-D scene understanding, but they rely on supervision from other tasks for joint learning. The second paradigm dedicates to developing fusion schemes to bridge two parallel modality streams. ACNet [8] proposes attention modules to exploit informative features for RGB-D semantic segmentation, whereas ABMDRNet [11] suggests reducing the modality differences of features before selectively extracting discriminative cues for RGB-T fusion. For RGB-P segmentation, Xiang *et al.* [12] connect RGB- and polarization branches via channel attention bridges. For RGB-E parsing, Zhang *et al.* [13] explore sparse-to-dense and dense-to-sparse fusion flows to extract dynamic context for accident scene segmentation. Salient object detection, seen as a specific type of image segmentation, can also benefit from multimodal fusion to identify the most important objects, such as Hyperfusion-Net [50] tailored for RGB-D and CAVER [51] for RGB-D and RGB-T. In this research, we also advocate this paradigm but unlike previous works, we address RGB-X semantic segmentation with a unified framework, for generalizing to diverse sensing modality combinations.

While previous works use a simple global channel-wise strategy, it does not work well across different sensing data. For example, ACNet [8] and SA-Gate [9], designed for RGB-D segmentation, perform less satisfactorily in RGB-T scene parsing [11]. In contrast, we hypothesize that comprehensive cross-modal interactions are crucial for RGB-X semantic segmentation with various supplements and uncertainties, so as to fully unleash the potential of cross-modal complementary features. Besides, most of the previous works adopt CNN backbone without considering that long-range dependency. We put forward a framework with transformers, which has global dependencies already in its architecture design. Differing from existing works, we perform fusion on different levels with cross-modal feature rectification and cross-attentional exchanging for enhanced dense semantic prediction.

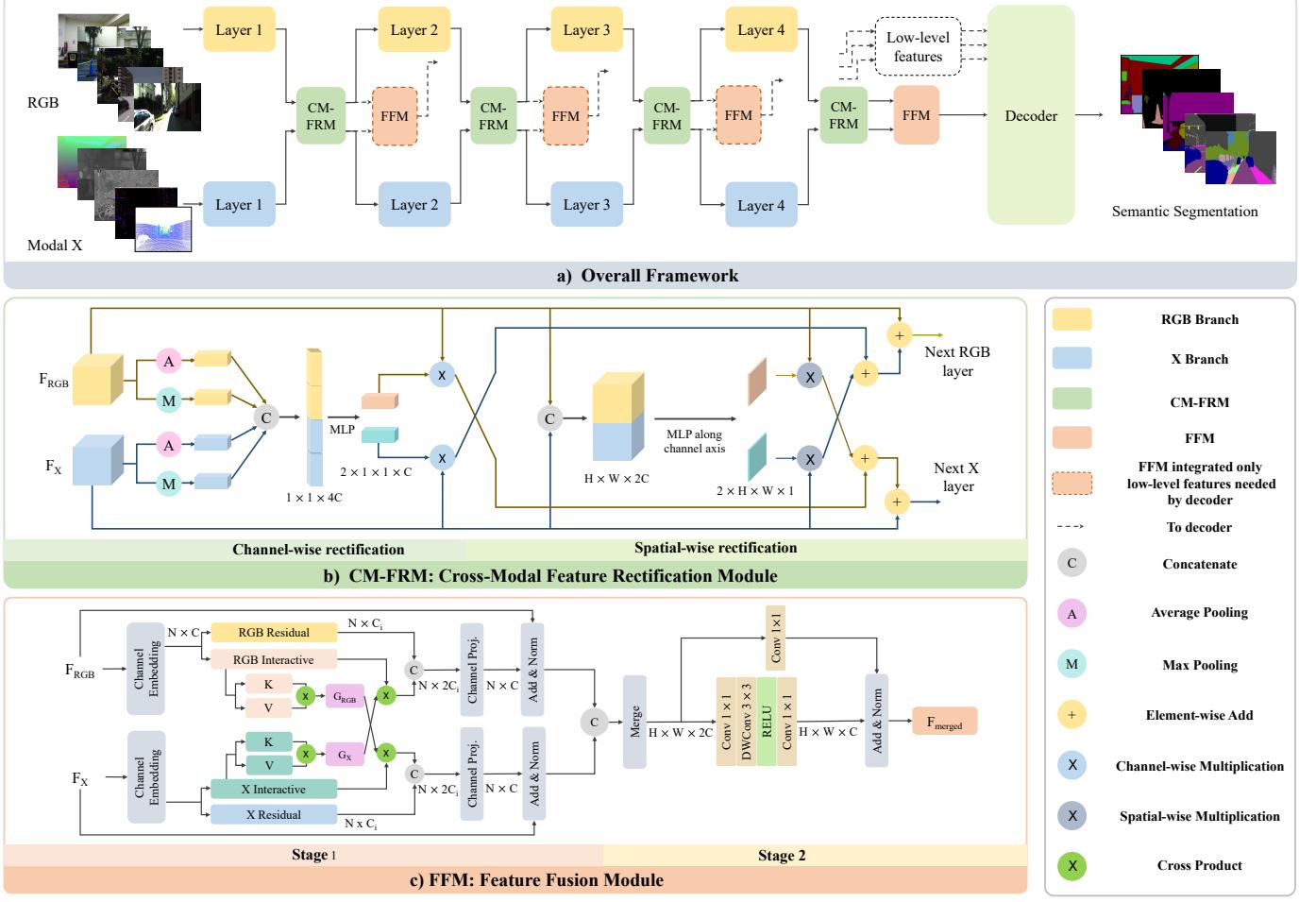


Fig. 4: a) Overview of CMX for *RGB-X semantic segmentation*. The inputs are RGB and another modality (*e.g.*, Depth, Thermal, Polarization, Event, or LiDAR). b) Cross-Modal Feature Rectification Module (*CM-FRM*) with colored arrows as information flows of the two modalities. c) Feature Fusion Module (*FFM*) with two stages of information exchange and fusion.

### III. PROPOSED FRAMEWORK: CMX

#### A. Framework Overview

The overview of CMX is shown in Fig. 4a. We use two parallel branches to extract features from RGB- and X-modal inputs, which can be RGB-Depth, -Thermal, -Polarization, -Event, -LiDAR data, *etc.* Specifically, our proposed framework for RGB-X semantic segmentation adopts a two-branch design to effectively extract features from both RGB and X modal inputs. The two branches involve the simultaneous processing of RGB and X modal data in a parallel but interactive manner, each of which is designed to capture the unique characteristics of the respective input modality. We introduce a rectification mechanism between both branches, enabling the feature from one modality to be rectified based on the feature from another modality. Additionally, we facilitate cross-modal feature interaction by exchanging rectified features from both modalities at each stage of the two-branch architecture. Based on two-branch architecture, our framework leverages the complementary information of both modalities to enhance the performance of RGB-X semantic segmentation.

While features from different modalities have their specific noisy measurements, the feature of another modality has the potential for rectifying and calibrating the noisy information. As shown in Fig. 4b, we design a Cross-Modal Feature Rectification Module (*CM-FRM*) to rectify one feature regarding another feature, and vice versa. In this manner, features from both modalities can be rectified. Besides, CM-FRMs are assembled between two adjacent stages of backbones. In this way, both rectified features are sent to the next stage to further deepen and improve the feature extraction. Furthermore, as shown in Fig. 4c, we design a two-stage Feature Fusion Module (*FFM*) to fuse features belonging to the same level into a single feature map. Then, a decoder is used to predict the final semantic map. In Sec. III-B and Sec. III-C, we detail the design of CM-FRM and FFM, respectively. In the following, we use X to refer to the supplementary modality, which can be Depth-, Thermal-, Polarization-, Event-, LiDAR data, *etc.*

#### B. Cross-Modal Feature Rectification

As analyzed above, the information originating from different sensing modalities are usually complementary [8], [9]

but contain noisy measurements. The noisy information can be filtered and calibrated by using features coming from another modality. To this purpose, in Fig. 4b, we propose a novel Cross-Modal Feature Rectification Module (*CM-FRM*) to perform feature rectification between parallel streams at each stage in feature extraction. To tackle noises and uncertainties in diverse modalities, CM-FRM processes features in two dimensions, including *channel-wise* and *spatial-wise* feature rectification, which together offer a holistic calibration, enabling better multi-modal feature extraction and interaction.

**Channel-wise feature rectification.** We embed bi-modal features  $\mathbf{RGB}_{in} \in \mathbb{R}^{H \times W \times C}$  and  $\mathbf{X}_{in} \in \mathbb{R}^{H \times W \times C}$  along the spatial axis into two attention vectors  $\mathbf{W}_{RGB}^C \in \mathbb{R}^C$  and  $\mathbf{W}_X^C \in \mathbb{R}^C$ . Different from previous channel-wise attention methods [9], [17], [52], we apply both global max pooling and global average pooling to  $\mathbf{RGB}_{in}$  and  $\mathbf{X}_{in}$  along the channel dimension to retain more information. We concatenate the four resulted vectors, having  $\mathbf{Y} \in \mathbb{R}^{4C}$ . Then, an MLP is applied, followed by a sigmoid function to obtain  $\mathbf{W}^C \in \mathbb{R}^{2C}$  from  $\mathbf{Y}$ , which will be split into  $\mathbf{W}_{RGB}^C$  and  $\mathbf{W}_X^C$ :

$$\mathbf{W}_{RGB}^C, \mathbf{W}_X^C = \mathcal{F}_{split}\left(\sigma\left(\mathcal{F}_{mlp}(\mathbf{Y})\right)\right), \quad (1)$$

where  $\sigma(\cdot)$  denotes the sigmoid function. The channel-wise rectification is then operated as:

$$\begin{aligned} \mathbf{RGB}_{rec}^C &= \mathbf{W}_X^C \circledast \mathbf{X}_{in}, \\ \mathbf{X}_{rec}^C &= \mathbf{W}_{RGB}^C \circledast \mathbf{RGB}_{in}, \end{aligned} \quad (2)$$

where  $\circledast$  denotes channel-wise multiplication.

**Spatial-wise feature rectification.** As the aforementioned channel-wise feature rectification module concentrates on learning global weights for a global calibration, we further introduce a spatial-wise feature rectification for calibrating local information. The bi-modal inputs  $\mathbf{RGB}_{in}$  and  $\mathbf{X}_{in}$  will be concatenated and embedded into two spatial weight maps:  $\mathbf{W}_{RGB}^S \in \mathbb{R}^{H \times W}$  and  $\mathbf{W}_X^S \in \mathbb{R}^{H \times W}$ . The embedding operation has two  $1 \times 1$  convolution layers assembled with a RELU function. Afterward, a sigmoid function is applied to obtain the embedded feature map  $\mathbf{F} \in \mathbb{R}^{H \times W \times 2}$ , which is further split into two weight maps. The process to obtain the spatial weight maps is formulated as:

$$\mathbf{F} = \text{Conv}_{1 \times 1}\left(\text{RELU}\left(\text{Conv}_{1 \times 1}(\mathbf{RGB}_{in} \parallel \mathbf{X}_{in})\right)\right), \quad (3)$$

$$\mathbf{W}_{RGB}^S, \mathbf{W}_X^S = \mathcal{F}_{split}\left(\sigma(\mathbf{F})\right). \quad (4)$$

Similar to channel-wise rectification, spatial-wise rectification is formulated as:

$$\begin{aligned} \mathbf{RGB}_{rec}^S &= \mathbf{W}_X^S * \mathbf{X}_{in}, \\ \mathbf{X}_{rec}^S &= \mathbf{W}_{RGB}^S * \mathbf{RGB}_{in}, \end{aligned} \quad (5)$$

where  $*$  denotes spatial-wise multiplication.

The whole rectified feature for both modalities  $\mathbf{RGB}_{out}$  and  $\mathbf{X}_{out}$  is organized as:

$$\begin{aligned} \mathbf{RGB}_{out} &= \mathbf{RGB}_{in} + \lambda_C \mathbf{RGB}_{rec}^C + \lambda_S \mathbf{RGB}_{rec}^S, \\ \mathbf{X}_{out} &= \mathbf{X}_{in} + \lambda_C \mathbf{X}_{rec}^C + \lambda_S \mathbf{X}_{rec}^S. \end{aligned} \quad (6)$$

$\lambda_C$  and  $\lambda_S$  are two hyperparameters. We set them both as 0.5 as default and will ablate in Sec. V-F.  $\mathbf{RGB}_{out}$  and  $\mathbf{X}_{out}$  are the rectified features after the comprehensive calibration, which will be sent into the next stage for feature fusion.

### C. Feature Fusion

After obtaining the feature maps at each layer, we build a two-stage Feature Fusion Module (*FFM*) to enhance the information interaction and combination. As shown in Fig. 4(c), in the information exchange stage (Stage 1), the two branches are still maintained, and a cross-attention mechanism is designed to globally exchange information between the two branches. In the fusion stage (Stage 2), the concatenated feature is transformed into the original size via a mixed channel embedding.

**Information exchange stage.** At this stage, the bi-modal features will exchange their information via a symmetric dual-path structure. For brevity, we take the X-modal path for illustration. We first flatten the input feature with size  $\mathbb{R}^{H \times W \times C}$  to  $\mathbb{R}^{N \times C}$ , where  $N = H \times W$ . Afterward, a linear embedding is used to generate two vectors with the same size  $\mathbb{R}^{N \times C_i}$ , which we call residual vector  $\mathbf{X}^{\text{res}}$  and interactive vector  $\mathbf{X}^{\text{inter}}$ . We further put forward an efficient cross-attention mechanism applied to these two interactive vectors from different modal paths, which will carry out sufficient information exchange across modalities. This offers complementary interactions from the sequence-to-sequence perspective beyond the rectification-based interactions from the feature map perspective in CM-FRM.

Our cross-attention mechanism for enhancing cross-modal feature fusion is based on the traditional self-attention [20]. The original self-attention operation encodes the input vectors into Query ( $\mathbf{Q}$ ), Key ( $\mathbf{K}$ ), and Value ( $\mathbf{V}$ ). The global attention map is calculated via a matrix multiplication  $\mathbf{Q}\mathbf{K}^T$ , which has a size of  $\mathbb{R}^{N \times N}$  and causes a high memory occupation. In contrast, [53] uses a global context vector  $\mathbf{G} = \mathbf{K}^T \mathbf{V}$  with a size  $\mathbb{R}^{C_{\text{head}} \times C_{\text{head}}}$  and the attention result is calculated by  $\mathbf{Q}\mathbf{G}$ . We flexibly adapt the reformulation and develop our multi-head cross-attention based on this efficient self-attention mechanism. Specifically, the interactive vectors will be embedded into  $\mathbf{K}$  and  $\mathbf{V}$  for each head, and both sizes of them are  $\mathbb{R}^{N \times C_{\text{head}}}$ . The output is obtained by multiplying the interactive vector and the context vector from the other modality path, namely a cross-attention process, and it is depicted in the following equations:

$$\begin{aligned} \mathbf{G}_{RGB} &= \mathbf{K}_{RGB}^T \mathbf{V}_{RGB}, \\ \mathbf{G}_X &= \mathbf{K}_X^T \mathbf{V}_X, \end{aligned} \quad (7)$$

$$\begin{aligned} \mathbf{U}_{RGB} &= \mathbf{X}_{RGB}^{\text{inter}} \text{SoftMax}(\mathbf{G}_X), \\ \mathbf{U}_X &= \mathbf{X}_X^{\text{inter}} \text{SoftMax}(\mathbf{G}_{RGB}). \end{aligned} \quad (8)$$

Note that  $\mathbf{G}$  denotes the global context vector, while  $\mathbf{U}$  indicates the attended result. To realize the attention from different representation subspaces, we remain the multi-head mechanism, where the number of heads matches the transformer backbone. Then, the attended result vector  $\mathbf{U}$  and the residual vector  $\mathbf{X}^{\text{res}}$  are concatenated. Finally, we apply a second linear embedding and resize the feature to  $\mathbb{R}^{H \times W \times C}$ .

**Fusion stage.** In the second stage of FFM, precisely the fusion stage, we use a simple channel embedding to merge the two paths’ features, which is realized via  $1 \times 1$  convolution layers. Further, we consider that during such a channel-wise fusion, the information of surrounding areas should also be exploited for robust RGB-X segmentation. Thereby, inspired by Mix-FFN in [33] and ConvMLP [54], we add one more depth-wise convolution layer  $DWConv_{3 \times 3}$  to realize a skip-connected structure. In this way, the merged features with the size  $\mathbb{R}^{H \times W \times 2C}$  are fused into the final output with the size of  $\mathbb{R}^{H \times W \times C}$  for feature decoding.

#### D. Multi-modal Data Representations

**RGB-Depth.** Depth images naturally offer range, position, and contour information. The fusion of RGB and depth information can better separate objects with indistinguishable colors and textures at different spatial locations. We encode the depth images into HHA format [55]. HHA offers geometric properties, including horizontal disparity, height above ground, and angle.

**RGB-Thermal.** At night or in places with insufficient light, objects, and backgrounds have similar color information and are difficult to distinguish. Thermal images provide infrared characteristics of objects, which are the potential to improve objects with thermal properties such as *people*. We directly use the infrared thermal image and copy the single-channel thermal image input 3 times to match the backbone input.

**RGB-Polarization.** High-reflectivity objects such as *glasses* and *cars* in RGB images are easily confused with surroundings. Polarization cameras record the optical polarimetric information when polarized reflection occurs, which offers complementary information in scenes with specular surfaces. The polarization sensor is equipped with a polarization mask layer with four different directions [12] and thereby each captured image set consists of four pixel-aligned images at different polarization angles  $[I_{0^\circ}, I_{45^\circ}, I_{90^\circ}, I_{135^\circ}]$ , where  $I_{angle}$  denotes the image recorded at the corresponding angle.

We investigate two representations, *i.e.*, the Degree of Linear Polarization (*DoLP*) and the Angle of Linear Polarization (*AoLP*), which are key polarimetric properties characterizing light polarization patterns [12]. They are derived by Stokes vectors  $S=\{S_0, S_1, S_2, S_3\}$  that describe the polarization state of light. Precisely,  $S_0$  represents the total light intensity,  $S_1$  and  $S_2$  denote the ratio of  $0^\circ$  and  $45^\circ$  linear polarization over its perpendicular polarized portion, and  $S_3$  stands for the circular polarization power which is not involved in our work. The Stokes vectors  $S_0, S_1, S_2$  can be calculated from image intensity measurements  $\{I_{0^\circ}, I_{45^\circ}, I_{90^\circ}, I_{135^\circ}\}$  via:

$$\begin{aligned} S_0 &= I_{0^\circ} + I_{90^\circ} = I_{45^\circ} + I_{135^\circ}, \\ S_1 &= I_{0^\circ} - I_{90^\circ}, \\ S_2 &= I_{45^\circ} - I_{135^\circ}. \end{aligned} \quad (9)$$

Then, *DoLP* and *AoLP* are formally computed as:

$$DoLP = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}, \quad (10)$$

$$AoLP = \frac{1}{2} \arctan \left( \frac{S_2}{S_1} \right). \quad (11)$$

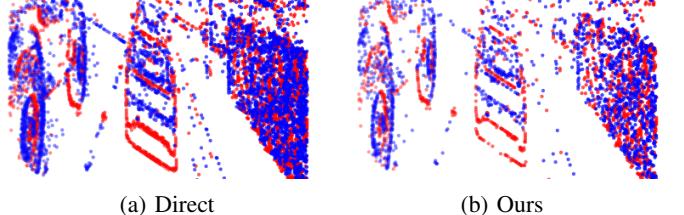


Fig. 5: Comparison between event representations.

In our experiments, we further study monochromatic and trichromatic polarization cues, coupled with RGB images in multi-modal RGB-P semantic segmentation. For monochromatic representation used in previous works [12], [56], we obtain it from monochromatic intensity measurements and convert it to 3-channel input by copying the single-channel information. For trichromatic polarization representation in either *DoLP* or *AoLP*, we compute separately for their respective RGB channels.

**RGB-Event.** Event data provide multiple advantages such as high dynamic range, high temporal resolution, and not being influenced by motion blur [57], which are critical in dynamic scenes with motion information such as road-driving environments [13], [44]. To process event data, a set of raw events in a time window  $\Delta T=t_N-t_1$  is embedded into a voxel grid with spatial dimensions  $H \times W$  and time bins  $B$ , where  $t_1$  and  $t_N$  are the start- and the end time stamp. Unlike previous work [26] converting event data to  $B=3$ , in this work, events are first embedded into a voxel grid with a higher time resolution, which we set the upscale size of the event bin as 6. Then, every 6 panels are superimposed to obtain a fine-grained event embedding. A comparison between the *direct* representation [26] and our event representation is shown in Fig. 5, in which our representation is more fine-grained in each event panel. Apart from  $B=3$ , we further investigate different settings of event time bin  $B=\{1, 5, 10, 15, 20, 30\}$  in our method for reaching robust RGB-E semantic segmentation.

**RGB-LiDAR.** LiDAR camera can provide reliable and accurate spatial-depth information on the physical world [14]. To make the representation of LiDAR data consistent with RGB images, we follow [14] to convert LiDAR data to a range-view image-like format. The Field-of-View (FoV) of the camera is  $90^\circ$  and the image resolution is  $H \times W=1408 \times 376$ . The origin is  $(u_0, v_0)=(H/2, W/2)$ . Then, the focal length  $(f_x, f_y)$  can be calculated through:

$$\begin{aligned} f_x &= H/(2 \times \tan(FoV \times \pi/360)), \\ f_y &= W/(2 \times \tan(FoV \times \pi/360)). \end{aligned} \quad (12)$$

Similar to [58], we project the LiDAR 3D points from the world coordinate to the 2D image coordinate by using:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}_{3 \times 1}^T & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \quad (13)$$

where  $(X, Y, Z)$  is the LiDAR point,  $(u, v)$  is the 2D image pixel, and the rotation  $(\mathbf{R})$  and the translation  $(\mathbf{t})$  matrices are given by KITTI-360 dataset [25].

## IV. EXPERIMENT DATASETS AND SETUPS

### A. Datasets

We use five RGB-Depth semantic segmentation datasets, and datasets of RGB-Thermal, RGB-Polarization, RGB-Event, and RGB-LiDAR combinations to verify our proposed CMX.

**NYU Depth V2 dataset** [24] contains 1449 RGB-D images with the size  $640 \times 480$ , divided into 795 training images and 654 testing images with annotations on 40 semantic categories.

**SUN-RGBD dataset** [59] has 10335 RGB-D images with 37 classes, and 5285/5050 for training/testing. Following [9], [60], we randomly crop and resize the input to  $480 \times 480$ .

**Stanford2D3D dataset** [61] has 70496 RGB-D images with 13 object categories. Following the data splitting [15], [45], areas of  $\{1, 2, 3, 4, 6\}$  are used for training and area 5 is for testing. The input image is resized to  $480 \times 480$ .

**ScanNetV2 dataset** [62] provides 19466/5436/2135 RGB-D samples for training/validation/testing. There are 20 classes. During training, the RGB images are re-scaled to the same size of  $640 \times 480$  as the depth images. During testing, the predictions are in the original size of  $1296 \times 968$ .

**Cityscapes dataset** [63] is an outdoor RGB-D dataset of urban road-driving street scenes. It is divided into 2975/500/1525 images in the training/validation/testing splits, both with finely annotated dense labels on 19 classes. The image scenes cover 50 different cities with a full resolution of  $2048 \times 1024$ .

**RGB-T MFNet dataset** [10] is a multi-spectral RGB-Thermal image dataset, which has 1569 images annotated in 8 classes at the resolution of  $640 \times 480$ . 820 images are captured during the day and the other 749 are at night. The training set has 50% of the daytime- and 50% of the nighttime images, while the validation- and test set respectively have 25% of the daytime- and 25% of the nighttime images.

**RGB-P ZJU dataset** [12] is an RGB-Polarization dataset collected by a multi-modal vision sensor designed for automated driving [18] on complex campus street scenes. It is composed of 344 images for training and 50 images for evaluation, both labeled with 8 semantic classes at the pixel level. The input image is resized to  $612 \times 512$ .

**RGB-E EventScape dataset.** A large-scale multi-modal RGB-Event semantic segmentation benchmark is not available. To fill this gap, we create an RGB-Event multi-modal semantic segmentation benchmark<sup>1</sup> based on the EventScape dataset [26], which is originally designed for depth estimation. The comparison between three event-based semantic segmentation datasets is presented in Table I. Unlike previous datasets using gray-scale images and pseudo labels, the RGB and the synthetic labels are available in our benchmark, which can provide more sufficient information and more precise annotations. To maintain data diversity from the original sequences generated by CARLA simulator [64], we select one frame from every 30 frames, obtaining 4077/749 images from 122329/22493 for training/evaluation. The images have a  $512 \times 256$  resolution and are annotated with 12 semantic classes, including Vehicle, Building, Wall, Vegetation, Road, Pole,

TABLE I: COMPARISON OF EVENT-BASED SEMANTIC SEGMENTATION DATASETS.

Dataset	Image	Event	Train/Val	Label	Resolution	Class
DDD17 [57]	Gray-scale	50Hz	15950/3890	pseudo	$346 \times 260$	6
DSEC-Semantic [65]	Gray-scale	20Hz	8082/2809	pseudo	$640 \times 440$	11
EventScape [26]	RGB	500Hz	122329/22493	synthetic	$512 \times 256$	12

RoadLines, Fences, Pedestrian, TrafficSign, Sidewalk, and TrafficLight.

**RGB-L KITTI-360 dataset.** KITTI-360 [25] is a suburban driving dataset, which has 49004/12276 images at the size of  $1408 \times 376$  for training/validation. There are 19 semantic classes following the Cityscapes dataset [63].

### B. Implementation Details

During training on all datasets, data augmentation is performed by random flipping and scaling with random scales  $[0.5, 1.75]$ . We take Mix Transformer encoder (MiT) pre-trained on ImageNet [66] as the backbone and MLP-decoder with an embedding dimension of 512 unless specified, both introduced in SegFormer [33]. We select AdamW optimizer [67] with weight decay 0.01. The original learning rate is set as  $6e^{-5}$  and we employ a poly learning rate schedule. We use cross-entropy as the loss function. When reporting multi-scale testing results on NYU Depth V2 and SUN RGB-D, we use multiple scales  $\{0.75, 1, 1.25\}$  with horizontal flipping. We use mean Intersection over Union (mIoU) averaged across semantic classes as the primary evaluation metric to measure the segmentation performance. More specific settings for different datasets are described in detail in the appendix.

## V. EXPERIMENTAL RESULTS AND ANALYSES

In this section, we present experimental results to verify the effectiveness of our proposed CMX for RGB-X semantic segmentation. In Sec. V-A, we show the results of CMX on multiple indoor and outdoor RGB-Depth benchmarks, compared with state-of-the-art methods. In Sec. V-B, we analyze the RGB-Thermal segmentation performance for robust daytime- and nighttime semantic perception. In Sec. V-C and Sec. V-D, we study the generalization of CMX to RGB-Polarization and RGB-Event modality combinations and representations of these multi-modal data. In Sec. V-E, we present the results of CMX on the RGB-LiDAR dataset. In Sec. V-F, we conduct a comprehensive variety of ablation studies to confirm the effects of different components in our solution. Finally, we perform efficiency- and qualitative analysis in Sec. V-G and Sec. V-H.

### A. Results on RGB-Depth Datasets

We first conduct experiments on RGB-D semantic segmentation datasets. The results are grouped in Table II.

**NYU Depth V2.** The results on the NYU Depth V2 dataset are shown in Table IIa. It can be easily seen that our approach achieves leading scores. The proposed method with MiT-B2 already exceeds previous methods, attaining 54.4% in mIoU. Our CMX models based on MiT-B4 and -B5 further

<sup>1</sup><https://paperswithcode.com/sota/semantic-segmentation-on-eventscape>

TABLE II: RESULTS ON FIVE RGB-DEPTH DATASETS. *Acc* AND \* DENOTE PIXEL ACCURACY AND MULTI-SCALE TEST.

(a) Results on NYU Depth V2 [24].

Method	mIoU (%)	Acc (%)
3DGNN [68]	43.1	-
Kong <i>et al.</i> [69]	44.5	72.1
LS-DeconvNet [70]	45.9	71.9
CFN [71]	47.7	-
ACNet [8]	48.3	-
RDF-101 [72]	49.1	75.6
SGNet [16]	51.1	76.8
ShapeConv [15]	51.3	76.4
NANet [60]	52.3	77.9
SA-Gate [9]	52.4	77.9
CMX (MiT-B2)	54.1	78.7
CMX (MiT-B2)*	<b>54.4</b>	<b>79.9</b>
CMX (MiT-B4)	56.0	79.6
CMX (MiT-B4)*	<b>56.3</b>	<b>79.9</b>
CMX (MiT-B5)	56.8	79.9
CMX (MiT-B5)*	<b>56.9</b>	<b>80.1</b>

(b) Results on Stanford2D3D [61].

Method	mIoU (%)	Acc (%)
Depth-aware CNN [45]	39.5	65.4
MMAF-Net-152 [73]	52.9	76.5
ShapeConv-101 [15]	60.6	<b>82.7</b>
CMX (MiT-B2)	61.2	82.3
CMX (MiT-B4)	<b>62.1</b>	82.6

(c) Results on SUN-RGBD [59].

Method	mIoU (%)	Acc (%)
3DGNN [68]	45.9	-
RDF-152 [72]	47.7	81.5
CFN [71]	48.1	-
D-CNN [45]	42.0	-
ACNet [8]	48.1	-
TCD [74]	49.5	83.1
SGNet [16]	48.6	82.0
SA-Gate [9]	49.4	82.5
NANet [60]	48.8	82.3
ShapeConv [15]	48.6	82.2
CMX (MiT-B2)*	49.7	82.8
CMX (MiT-B4)*	52.1	83.5
CMX (MiT-B5)*	<b>52.4</b>	<b>83.8</b>

(d) Results on ScanNetV2 *test* set [62].

Method	Modal	mIoU (%)
PSPNet [6]	RGB	47.5
AdapNet++ [75]	RGB	50.3
3DMV (2d-proj) [76]	RGB-D	49.8
FuseNet [77]	RGB-D	53.5
SSMA [75]	RGB-D	57.7
GRBNet [38]	RGB-D	59.2
MCA-Net [78]	RGB-D	59.5
DMMF [79]	RGB-D	59.7
CMX (MiT-B2)	RGB-D	<b>61.3</b>

(e) Results on Cityscapes *val* set [63].

Method	Modal	Backbone	mIoU (%)
SwiftNet [80]	RGB	ResNet-18	70.4
ESANet [81]	RGB	ResNet-50	79.2
GSCNN [82]	RGB	WideResNet-38	80.8
CCNet [30]	RGB	ResNet-101	81.3
DANet [7]	RGB	ResNet-101	81.5
ACFNet [83]	RGB	ResNet-101	81.5
SegFormer [33]	RGB	MiT-B2	81.0
SegFormer [33]	RGB	MiT-B4	82.3
RFNet [3]	RGB-D	ResNet-18	72.5
PADNet [84]	RGB-D	ResNet-50	76.1
Kong <i>et al.</i> [69]	RGB-D	ResNet-101	79.1
ESANet [81]	RGB-D	ResNet-50	80.0
SA-Gate [9]	RGB-D	ResNet-50	80.7
SA-Gate [9]	RGB-D	ResNet-101	81.7
AsymFusion [85]	RGB-D	Xception65	82.1
SSMA [75]	RGB-D	ResNet-50	82.2
CMX	RGB-D	MiT-B2	<b>81.6</b>
CMX	RGB-D	MiT-B4	<b>82.6</b>

dramatically improve the mIoU to 56.3% and 56.9%, clearly standing out in front of all state-of-the-art approaches. The best CMX model even reaches superior results than recent strong pretraining-based methods [19], [49] like Omnivore [19] that uses images, videos, and single-view 3D data for supervision.

**Stanford2D3D.** In Table IIb, our CMX achieves state-of-the-art mIoU scores. Our B2-based CMX surpasses the previous best ShapeConv [15] based on ResNet-101 [86] in mIoU and our model based on MiT-B4 further reaches mIoU of 62.1%. The results demonstrate the effectiveness and learning capacity of our approach on such a large RGB-D dataset.

**SUN-RGBD.** As presented in Table IIC, our method achieves leading performances on the SUN-RGBD dataset. Our interactive cross-modal fusion approach (Fig. 2c) exceeds previous input fusion methods (Fig. 2a), *e.g.*, SGNet [16] and ShapeConv [15], as well as feature fusion methods (Fig. 2b), *e.g.*, ACNet [8] and SA-Gate [9]. In particular, with MiT-B4 and -B5, CMX elevates the mIoU to >52.0%. CMX is also better than multi-task methods like PAP [48] and TET [87].

**ScanNetV2.** We test our CMX model with MiT-B2 on the ScanNetV2 benchmark. As shown in Table IID, it can be clearly seen that CMX outperforms RGB-only methods and achieves the top mIoU of 61.3% among the RGB-D methods. On the ScanNetV2 leaderboard, methods like BPNet [88] reach higher scores by using 3D supervision from point clouds to perform joint 2D- and 3D reasoning. In contrast, our method attains a competitively accurate performance by using purely 2D data and effectively leveraging the complementary information inside RGB-D modalities.

**Cityscapes.** Besides indoor RGB-D datasets, to study the generalizability to outdoor scenes, we assess the effectiveness of CMX on Cityscapes. As shown in Table IIE, we note that the improvement on the Cityscapes dataset is not as obvious as

other datasets, because the performance of RGB-only models on this dataset shows a saturation trend. Compared with MiT-B2 (RGB), our RGB-D approach elevates the mIoU by 0.6%. Our approach based on MiT-B4 achieves a state-of-the-art score of 82.6%, outstripping all existing RGB-D methods by more than 0.4% in absolute mIoU values, verifying that CMX generalizes well to street scene understanding.

### B. Results on RGB-Thermal Dataset

**Comparison with the state-of-the-art.** In Table III, we compare our method against RGB-only models and multi-modal methods using RGB-T inputs of MFNet dataset [10]. As unfolded, ACNet [8] and SA-Gate [9], carefully designed for RGB-Depth segmentation, perform less satisfactorily on RGB-T data, as they focus on feature extraction without sufficient feature interaction before fusion and thereby fail to generalize to other modality. Depth-aware CNN [45], an input fusion method with modality-specific operator design, also does not yield high performance. In contrast, the proposed CMX strategy, enabling comprehensive interactions from various perspectives, generalizes smoothly in RGB-T semantic segmentation. It can be seen that our method based on MiT-B2 achieves mIoU of 58.2%, clearly outperforming the previous best RGB-T methods ABMDRNet [11], FEANet [17], and GMNet [42]. Our CMX with MiT-B4 further elevates state-of-the-art mIoU to 59.7%, widening the accuracy gap in contrast to existing methods. Moreover, it is worth pointing out that the improvements brought by our RGB-X approach compared with the RGB-only baselines are compelling, *i.e.*, +5.0% and +4.9% in mIoU for MiT-B2 and -B4 backbones, respectively. Our approach overall achieves top scores on *car*, *person*, *bike*, *curve*, *car stop*, and *bump*. For *person* with infrared properties, our approach enjoys more than +11.0%

TABLE III: PER-CLASS RESULTS ON MFNET DATASET [10] FOR RGB-THERMAL SEGMENTATION.

Method	Modal	Unlabeled	Car	Person	Bike	Curve	Car Stop	Guardrail	Color Cone	Bump	mIoU
ERFNet [89]	RGB	96.7	67.1	56.2	34.3	30.6	9.4	0.0	0.1	30.5	36.1
DANet [7]	RGB	96.3	71.3	48.1	51.8	30.2	18.2	0.7	30.3	18.8	41.3
PSPNet [6]	RGB	96.8	74.8	61.3	50.2	38.4	15.8	0.0	33.2	44.4	46.1
HRNet [90]	RGB	98.0	86.9	67.3	59.2	35.3	23.1	1.7	46.6	47.3	51.7
SegFormer-B2 [33]	RGB	97.9	87.4	62.8	63.2	31.7	25.6	9.8	50.9	49.6	53.2
SegFormer-B4 [33]	RGB	98.0	88.9	64.0	62.8	38.1	25.9	6.9	50.8	57.7	54.8
MFNet [10]	RGB-T	96.9	65.9	58.9	42.9	29.9	9.9	0.0	25.2	27.7	39.7
SA-Gate [9]	RGB-T	96.8	73.8	59.2	51.3	38.4	19.3	0.0	24.5	48.8	45.8
Depth-aware CNN [45]	RGB-T	96.9	77.0	53.4	56.5	30.9	29.3	8.5	30.1	32.3	46.1
ACNet [8]	RGB-T	96.7	79.4	64.7	52.7	32.9	28.4	0.8	16.9	44.4	46.3
PSTNet [91]	RGB-T	97.0	76.8	52.6	55.3	29.6	25.1	<b>15.1</b>	39.4	45.0	48.4
RTFNet [40]	RGB-T	98.5	87.4	70.3	62.7	45.3	29.8	0.0	29.1	55.7	53.2
FuseSeg [41]	RGB-T	97.6	87.9	71.7	64.6	44.8	22.7	6.4	46.9	47.9	54.5
AFNet [92]	RGB-T	98.0	86.0	67.4	62.0	43.0	28.9	4.6	44.9	56.6	54.6
ABMDRNet [11]	RGB-T	<b>98.6</b>	84.8	69.6	60.3	45.1	33.1	5.1	47.4	50.0	54.8
FEANet [17]	RGB-T	98.3	87.8	71.1	61.1	46.5	22.1	6.6	<b>55.3</b>	48.9	55.3
DHFNet [93]	RGB-T	97.7	87.6	71.7	61.1	39.5	<b>42.4</b>	9.5	49.3	56.0	57.2
GMNet [42]	RGB-T	97.5	86.5	73.1	61.7	44.0	42.3	14.5	48.7	47.4	57.3
CMX (MiT-B2)	RGB-T	98.3	89.4	74.8	<b>64.7</b>	47.3	30.1	8.1	52.4	59.4	58.2
CMX (MiT-B4)	RGB-T	98.3	<b>90.1</b>	<b>75.2</b>	64.5	<b>50.2</b>	35.3	8.5	54.2	<b>60.6</b>	<b>59.7</b>

TABLE IV: SEGMENTATION RESULTS ON DAYTIME- AND NIGHTTIME IMAGES ON MFNET DATASET [10].

Method	Modal	Daytime mIoU (%)	Nighttime mIoU (%)
FRRN [94]	RGB	40.0	37.3
DFN [95]	RGB	38.0	42.3
BiSeNet [96]	RGB	44.8	47.7
SegFormer-B2 [33]	RGB	48.6	49.2
SegFormer-B4 [33]	RGB	49.4	52.4
MFNet [10]	RGB-T	36.1	36.8
FuseNet [77]	RGB-T	41.0	43.9
RTFNet [40]	RGB-T	45.8	54.8
FuseSeg [41]	RGB-T	47.8	54.6
GMNet [42]	RGB-T	49.0	57.7
CMX (MiT-B2)	RGB-T	51.3	57.8
CMX (MiT-B4)	RGB-T	<b>52.5</b>	<b>59.4</b>

gain in IoU, confirming the effectiveness of CMX in harvesting complementary cross-modal information.

**Day and night performances.** Following [41], [42], we assess day- and night segmentation results on the RGB-T benchmark (see Table IV). For daytime scenes, our approach increases mIoU by 2.7%~3.1% compared with RGB-only baselines. At nighttime, RGB segmentation often suffers from poor lighting conditions, and it even carries much noisy information in the RGB data. Yet, our CMX rectifies the noisy images and exploits supplementary features from thermal data, dramatically improving the mIoU by >7.0% and enhancing the robustness of semantic scene understanding in unfavorable environments with adverse illuminations.

### C. Results on RGB-Polarization Dataset

**Comparison with the state-of-the-art.** Table V shows per-class accuracy of our approach compared to RGB-only [33], [80] and RGB-Polarization fusion methods [12], [56] on ZJU-RGB-P dataset [12]. Our unified CMX outperforms the previous best RGB-P method [12] by >6.0% in mIoU. We observe that the improvement on *pedestrian* is significant thanks to

the capacity of the transformer backbone and our cross-modal fusion mechanisms. Compared to the RGB-only baseline with MiT-B2 [33]), the IoU improvements on classes with polarimetric characteristics are clear, such as *glass* (>8.0%) and *car* (>2.5%), further evidencing the generalizability of our cross-modal fusion solution in bridging RGB-P streams.

**Analysis of polarization data representations.** We study polarimetric data representations and the results displayed in Table V indicate that the Angle of Linear Polarization (*AoLP*) and the Degree of Linear Polarization (*DoLP*) representations both carry effective polarization information beneficial for semantic scene understanding, which is consistent with the finding in [12]. Besides, trichromatic representations are consistently better than monochromatic representations used in previous RGB-P segmentation works [12], [56]. This is expected as the trichromatic representation provides more detailed information, which should be leveraged to fully unlock the potential of trichromatic polarization cameras.

### D. Results on RGB-Event Dataset

**Comparison with the state-of-the-art.** In Table VI, we benchmark more than 10 semantic segmentation methods, including RGB-only methods, CNN-based [80], [97], [98], [100] and transformer-based [23], [33], [99] methods, as well as multi-modal methods [3], [9], [13]. In contrast, our models improve performance by mixing RGB-Event features, as seen in Table VI and Fig. 6. Our model using MiT-B4 reaches 64.28% in mIoU, towering over all other methods and setting the state-of-the-art on the RGB-E benchmark. This further verifies the versatility of our solution for different multi-modal combinations. Fig. 6 depicts a per-class accuracy comparison between the RGB baseline and our RGB-Event model with MiT-B2. With event data, the foreground objects are more accurately parsed by our RGB-E model, e.g., *vehicle* (+2.1%), *pedestrian* (+11.7%), and traffic light (+7.0%).

TABLE V: PER-CLASS RESULTS ON ZJU-RGB-P [12] DATASET FOR RGB-POLARIZATION SEGMENTATION.

Method	Modal	Building	Glass	Car	Road	Vegetation	Sky	Pedestrian	Bicycle	mIoU
SwiftNet [80]	RGB	83.0	73.4	91.6	96.7	94.5	84.7	36.1	82.5	80.3
SegFormer-B2 [33]	RGB	90.6	79.0	92.8	96.6	96.2	89.6	82.9	89.3	89.6
NLFNet [56]	RGB-P	85.4	77.1	93.5	97.7	93.2	85.9	56.9	85.5	84.4
EAFNet [12]	RGB-P	87.0	79.3	93.6	97.4	95.3	87.1	60.4	85.6	85.7
CMX (SegFormer-B2)	RGB-AoLP (Monochromatic)	<b>91.9</b>	87.0	95.6	98.2	96.7	89.0	84.9	92.0	91.8
CMX (SegFormer-B2)	RGB-AoLP (Trichromatic)	91.5	87.3	95.8	98.2	96.6	89.3	85.6	91.9	92.0
CMX (SegFormer-B4)	RGB-AoLP (Monochromatic)	91.8	<b>88.8</b>	<b>96.3</b>	<b>98.3</b>	96.7	89.1	86.3	92.3	92.4
CMX (SegFormer-B4)	RGB-AoLP (Trichromatic)	91.6	<b>88.8</b>	<b>96.3</b>	<b>98.3</b>	<b>96.8</b>	89.7	86.2	<b>92.8</b>	<b>92.6</b>
CMX (SegFormer-B2)	RGB-DoLP (Monochromatic)	91.4	87.6	96.0	98.2	96.6	89.1	87.1	92.3	92.1
CMX (SegFormer-B2)	RGB-DoLP (Trichromatic)	91.8	87.8	96.1	98.2	96.7	<b>89.4</b>	86.1	91.8	92.2
CMX (SegFormer-B4)	RGB-DoLP (Monochromatic)	91.8	88.6	<b>96.3</b>	<b>98.3</b>	96.7	89.4	86.0	92.1	92.4
CMX (SegFormer-B4)	RGB-DoLP (Trichromatic)	91.6	88.6	<b>96.3</b>	<b>98.3</b>	96.7	89.5	<b>86.4</b>	92.2	92.5

TABLE VI: RESULTS FOR RGB-EVENT SEGMENTATION.

Method	Modal	Backbone	mIoU (%)	Pixel Acc. (%)
SwiftNet [80]	RGB	ResNet-18	36.67	83.46
Fast-SCNN [97]	RGB	Fast-SCNN	44.27	87.10
CGNet [98]	RGB	M3N21	44.75	87.13
Trans4Trans [99]	RGB	PVT-B2	51.86	89.03
Swin-s [23]	RGB	Swin-s	52.49	88.78
Swin-b [23]	RGB	Swin-b	53.31	89.21
DeepLabV3+ [100]	RGB	ResNet-101	53.65	89.92
SegFormer-B2 [33]	RGB	MiT-B2	58.69	91.21
SegFormer-B4 [33]	RGB	MiT-B4	59.86	91.61
RFNet [3]	RGB-E	ResNet-18	41.34	86.25
ISSAFE [13]	RGB-E	ResNet-18	43.61	86.83
SA-Gate [9]	RGB-E	ResNet-101	53.94	90.03
CMX (DeepLabV3+)	RGB-E	ResNet-101	54.91	89.67
CMX (Swin-s)	RGB-E	Swin-s	60.86	91.25
CMX (Swin-b)	RGB-E	Swin-b	61.21	91.61
CMX (SegFormer-B2)	RGB-E	MiT-B2	61.90	91.88
CMX (SegFormer-B4)	RGB-E	MiT-B4	<b>64.28</b>	<b>92.60</b>

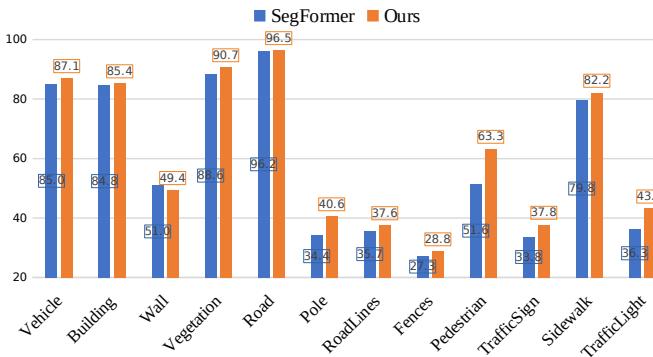


Fig. 6: Per-class IoU results of the RGB-only baseline and our RGB-Event model on our RGB-Event benchmark.

**Analysis of using different backbones.** To verify that our unified method is effective with using different backbones, we compare CNN- and transformer-based backbones in the CMX framework. Specifically, in addition to MiT backbones, we experiment with DeepLabV3+ [100] and Swin transformer [23] backbones with UperNet [101] to construct CMX. Compared to the RGB-only DeepLabV3+, Swin-s, and Swin-b methods, CMX models achieve respective +1.26%, +8.37%,

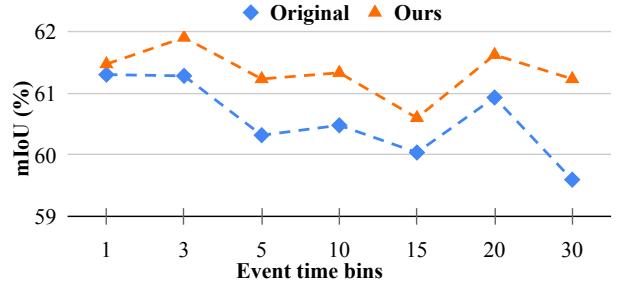


Fig. 7: Analysis of event representations and time bins.

+7.90% gains in mIoU. The results show that our RGB-X solution consistently improves the segmentation performance, confirming that our unified framework is not strictly tied to a concrete backbone type, but can be flexibly deployed with CNN- or transformer models, which helps to yield effective unified architecture for RGB-X semantic segmentation.

**Analysis of event data representations.** We study with different settings of event time bin  $B=\{1, 3, 5, 10, 15, 20, 30\}$  based on our CMX fusion model with MiT-B2. Compared with the original event representation [26], our representation achieves consistent improvements (in Fig. 7) on different settings of event time bins, such as +1.63% of mIoU when  $B=30$ . In particular, it helps our CMX to obtain the highest mIoU of 61.90% in the setting of  $B=3$ . In  $B=1$ , embedding all events in a single time bin leads to dragging behind images of moving objects and being sub-optimal for feature fusion. In higher time bins, events produced in a short interval are dispersed to more bins, resulting in insufficient events in a single bin. These corroborate observations in [13], [44] and that the event representation  $B=3$  is an effective time bin setting for RGB-E semantic segmentation with CMX.

#### E. Results on RGB-LiDAR Dataset

In Table VII, we compare CMX with other models dedicated to RGB-LiDAR data fusion, including PMF [14] and TransFuser [104]. These two methods achieve respective 54.48% and 56.57% in mIoU. Besides, other general multimodal fusion methods, e.g., HRFuser [102] and TokenFusion [103], are included for comparison. In contrast, our CMX obtains the

TABLE VII: RESULTS FOR RGB-LiDAR SEGMENTATION.

Method	Backbone	mIoU (%)
HRFuser [102]	HRFormer-T	48.74
PMF [14]	SalsaNext	54.48
TokenFusion [103]	MiT-B2	54.55
TransFuser [104]	RegNetY-3.2GF	56.57
CMX	MiT-B2	<b>64.31</b>

state-of-the-art performance with 64.31% in mIoU, having a +9.76% gain compared with TokenFusion which is also based on MiT-B2. The sufficient improvement proves the advantage of using a symmetric dual-stream architecture in modal fusion and the effectiveness of our proposed cross-modal rectification and fusion methods.

#### F. Ablation Study

We perform a series of ablation studies to explore how different parts of our architecture affect the segmentation. We use depth information encoded into HHA as the complementary modality here. We take MiT-B2 as the backbone with the MLP decoder in our ablation studies unless specified. The semantic segmentation performance is evaluated on NYU Depth V2.

**RGB-only Baseline and CMX.** In order to comprehensively compare the RGB-only baseline [33] and our RGB-X-based model, we conduct experiments on five different types of modality fusion, including RGB-Depth, -Thermal, -Polarization, -Event, and -LiDAR. Both methods are based on the same backbone with MiT-B2 [33]. As presented in Table VIII, on six different datasets, *i.e.*, NYU Depth V2, Cityscapes, MFNet, ZJU-RGB-P, EventScape, and KITTI-360, our CMX model obtains improvements of +6.1%, +0.6%, +5.0%, +2.6%, +3.2%, and +3.0%, respectively. We note that the improvement on the Cityscapes dataset is not as obvious as other datasets, because the performance of RGB-only models on this dataset shows a saturation trend. Nonetheless, the consistent improvements achieved across five different multi-modal fusion tasks are a strong testament to the effectiveness of our proposed unified CMX framework for RGB-X semantic segmentation.

**Effectiveness of CM-FRM and FFM.** We design CM-FRM and FFM to rectify and merge features coming from the RGB- and X-modality branches. We take out these two modules from the architecture respectively, where the results are shown in Table IX. If CM-FRM is ablated, the features will be extracted independently in their own branches, and for FFM we simply average the two features for semantic prediction. Compared with the baseline, using only CM-FRM improves mIoU by 2.5%, using only FFM improves mIoU by 1.2%, and together CM-FRM and FFM improve the semantic segmentation performance by 3.8%. The improvements show that our CM-FRM and FFM modules are both crucial for the success of the unified CMX framework.

**Ablation with CM-FRM and FFM variants.** We further experiment with variants of CM-FRM and FFM modules. As shown in Table X, *channel only* denotes using channel-wise rectification only ( $\lambda_C=1$  and  $\lambda_S=0$  in Eq. 6), and

*spatial only* means using spatial-wise rectification only ( $\lambda_C=0$  and  $\lambda_S=1$  in Eq. 6). It can be seen that substituting the proposed CM-FRM by either *channel-only* or *spatial-only* variant causes a sub-optimal accuracy, further confirming the efficacy of combining the bi-modal rectification for holistic feature calibration, which is crucial for robust multi-modal segmentation. In our channel-wise calibration, we use both global average pooling and global max pooling to retain more information. Table X shows that using only global average pooling (*avg. p.*) and using only global max pooling (*max. p.*) are less effective than our complete CM-FRM, which offers a more comprehensive rectification.

Previous ablation studies support the design of CM-FRM. To understand the capability of FFM, we here test with two variants. As shown in Table X, *stage 2 only* means there is no information exchange before the mixed channel embedding, whereas *self attn* denotes that context vectors will not be exchanged in stage 1 of FFM. The two variants are less constructive as compared to our complete FFM. Thanks to the crucial cross-attention design for information exchange, our complete FFM productively rectifies and fuses the features at different levels. These indicate the importance of fusion from the sequence-to-sequence perspective, which is not considered in previous works. Overall, the ablation shows that our interactive strategy, providing comprehensive interactions, is effective for cross-modal fusion.

**Ablation of the supplementary modality.** Previous works have shown that multi-modal segmentation has a better performance than single-modal RGB segmentation [8]. We carry out experiments to certify that and the results are shown in Table XI. Note that here, the MLP decoder is not used, in order to focus on studying the influence of feature extraction from different supplementary modalities. As compared to the RGB-only method, we conduct experiments with modalities of RGB-RGB, RGB-Noise, RGB-Depth, and RGB-HHA. We found that replacing the supplementary modality with random noise can obtain even better results than two RGB inputs. This means that even pure noise information may help the model identify noisy information in the RGB branch. The model learns to focus on relevant features and thus gains robustness. It may also help prevent over-fitting during the learning process. However, when using depth information, we have observed obvious improvements, which further proves that the fusion of RGB and depth information brings clearly better predictions. Encoding depth images using the HHA representation further increases the scores. The overall gain of 5.3% in mIoU, compared with the RGB-only baseline, is also compelling, which is similar to that in RGB-T semantic segmentation, demonstrating the effectiveness of our proposed method for rectifying and fusing cross-modal information.

#### G. Efficiency Analysis

In Table XII, we present the computational complexity results. Compared with the previous best method SA-Gate [9] on the NYU Depth V2 dataset, our model with MiT-B2 has similar #Params and lower FLOPs but significantly higher mIoU. Our CMX model with MiT-B4 greatly elevates the

TABLE VIII: COMPARISON BETWEEN RGB-ONLY BASELINE AND OUR CMX MODEL FOR RGB-X SEMANTIC SEGMENTATION, WHERE ALL RESULTS (MIoU) ARE BASED ON THE SAME BACKBONE WITH MiT-B2.

Method	Modal	NYU Depth V2	Cityscapes	MFNet	ZJU-RGB-P	EventScape	KITTI-360
SegFormer-B2 [33]	RGB-only	48.0	81.0	53.2	89.6	58.7	61.3
CMX-B2	Multimodal	54.1 (RGB-D)	81.6 (RGB-D)	58.2 (RGB-T)	92.2 (RGB-P)	61.9 (RGB-E)	64.3 (RGB-L)

TABLE IX: ABLATION STUDY OF CM-FRM AND FFM ON NYU DEPTH V2 *test* SET. Avg. IS THE AVERAGE FUSION.

CM-FRM	FFM	mIoU (%)	Pixel Acc. (%)
✗	Avg.	50.3	76.8
✓	Avg.	52.8	78.0
✗	✓	51.5	77.1
✓	✓	<b>54.1</b>	<b>78.7</b>

TABLE X: ABLATION WITH CM-FRM/FFM VARIANTS ON NYU DEPTH V2 *test* SET.

Feature Rectify	Feature Fusion	mIoU (%)	Pixel Acc. (%)
CM-FRM <i>channel only</i>	FFM	53.6	78.5
CM-FRM <i>spatial only</i>	FFM	53.3	78.3
CM-FRM <i>avg. p. only</i>	FFM	53.0	78.1
CM-FRM <i>max. p. only</i>	FFM	53.5	78.5
CM-FRM	<i>FFM stage 2 only</i>	53.8	78.5
CM-FRM	<i>FFM self attn</i>	53.8	78.6
CM-FRM	FFM	<b>54.1</b>	<b>78.7</b>

mIoU score to 56.0%, further widening the accuracy gap with moderate model complexity. With MiT-B5, mIoU further increases to 56.8%, but it also comes with larger complexity. For efficiency-critical applications, the CMX solution with MiT-B2 or -B4 would be preferred to enable both accurate and efficient multi-modal semantic scene perception.

#### H. Qualitative Analysis

**Visualization of segmentation results.** We compare the results of the RGB-only baseline and our CMX, where both are based on SegFormer-B2. We analyze each row from top to bottom in Fig. 8.

- (1) For RGB-Depth, we present results from the NYU Depth V2 dataset [24]. CMX leverages geometric information and correctly identifies the *bed* while the model wrongly classifies it as a *sofa*. It proves that the CMX model can obtain discriminative features from depth information in the low-texture scenario.
- (2) For RGB-Thermal, our CMX demonstrates improvement over the baseline under low illumination conditions, *e.g.*, the night scene. The use of Thermal in addition to RGB enables the model to make much clearer boundaries, such as between *persons* and *unlabeled* background. Besides, by combining features from both modalities, our CMX can more effectively filter out the noise and other unwanted artifacts that can negatively impact segmentation accuracy. For example, the segmentation of *persons* in

TABLE XI: ABLATION OF THE SUPPLEMENTARY MODALITY ON NYU DEPTH V2 *test* SET.

Modalities	mIoU (%)	Pixel Acc. (%)
RGB	46.7	73.8
RGB + RGB	47.2	74.1
RGB + Noise	47.7	74.5
RGB + Raw depth	<b>51.1</b>	<b>75.7</b>
RGB + HHA	<b>52.0</b>	<b>77.0</b>

TABLE XII: EFFICIENCY RESULTS. FLOPS ARE ESTIMATED FOR INPUTS OF RGB AND HHA, WITH A SIZE OF  $480 \times 640 \times 3$ .

Method	#Params (M)	FLOPs (G)	mIoU (%)
SA-Gate [9] (ResNet50)	<b>63.4</b>	204.9	50.4
CMX (SegFormer-B2)	66.6	<b>67.6</b>	54.1
CMX (SegFormer-B4)	139.9	134.3	56.0
CMX (SegFormer-B5)	181.1	167.8	<b>56.8</b>

the distance is easily disturbed by overexposed lights in RGB, which can be rectified by Thermal modality.

- (3) For RGB-Polarization, the specular *glass* areas are more precisely parsed by our CMX model, as compared to the baseline. Besides, the *cars* which also contain polarization cues are completely and smoothly segmented with delineated borders, and the boundaries of *pedestrians* also show beneficial effects.
- (4) For RGB-Event, our CMX generalizes well and enhances the segmentation of moving objects, such as the segmentation results of *cyclists* and *poles*. It indicates that incorporating features extracted from Event data can enhance the modeling of dynamics that are not captured by RGB images alone.
- (5) For RGB-LiDAR, thanks to the spatial information from the LiDAR modality, our CMX model can correctly recognize the *wall*, while the RGB-only method misidentifies it as part of a *truck*. Furthermore, our CMX module makes CMX robust against the noise of LiDAR modality, such as the *truck* glass area, yielding a complete segmentation mask of the *truck*.

Overall, the qualitative examination backs up that our general approach is suitable for a diverse mix of multi-modal sensing combinations for robust semantic scene understanding.

## VI. CONCLUSION

To revitalize multi-modal pixel-wise semantic scene understanding for autonomous vehicles, we investigate RGB-X semantic segmentation and propose CMX, a universal

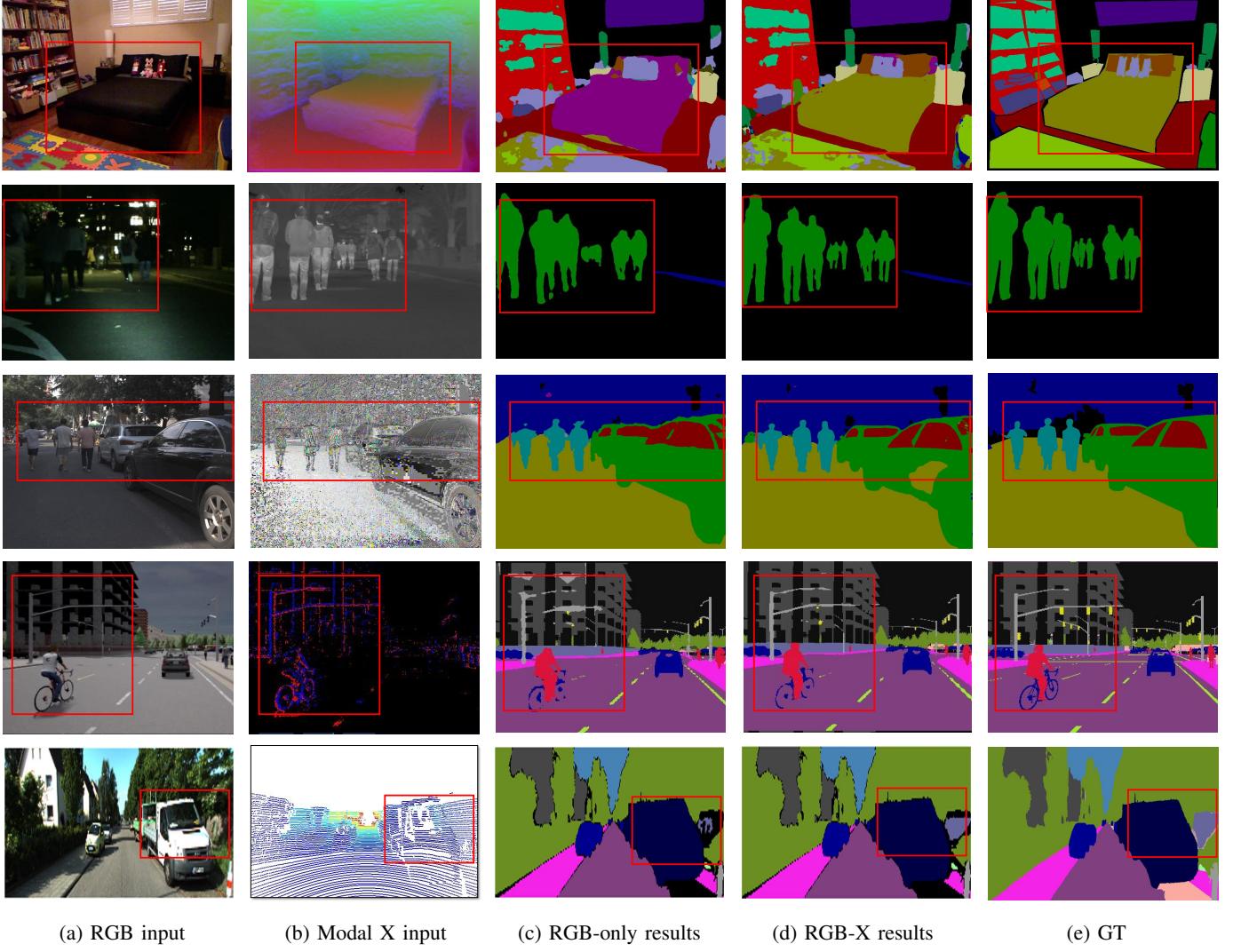


Fig. 8: Visualization results of RGB-only and RGB-X methods, where both are based on the same backbone. From top to bottom: RGB-Depth, RGB-Thermal, RGB-Polarization (AoLP), RGB-Event, and RGB-LiDAR semantic segmentation.

transformer-based cross-modal fusion architecture, which is generalizable to a diverse mix of sensing data combinations. We put forward a Cross-Modal Feature Rectification Module (CM-FRM) and a Feature Fusion Module (FFM) for facilitating interactions toward accurate RGB-X semantic segmentation. CM-FRM conducts channel- and spatial-wise rectification, rendering comprehensive feature calibration. FFM intertwines cross-attention and mixed channel embedding for enhanced global information exchange. To further assess the generalizability of CMX to dense-sparse data fusion, we establish an RGB-Event semantic segmentation benchmark. We study effective representations of polarimetric- and event data, indicating the optimal path to follow for reaching robust multi-modal semantic segmentation. The proposed model sets the new state-of-the-art on nine benchmarks, spanning five RGB-D datasets, as well as RGB-Thermal, RGB-Polarization, RGB-Event, and RGB-LiDAR combinations.

## REFERENCES

- [1] W. Zhou, J. S. Berrio, S. Worrall, and E. Nebot, “Automated evaluation of semantic segmentation robustness for autonomous driving,” *T-ITS*, vol. 21, no. 5, pp. 1951–1963, 2020.
- [2] K. Yang, X. Hu, Y. Fang, K. Wang, and R. Stiefelhagen, “Omnisupervised omnidirectional semantic segmentation,” *T-ITS*, vol. 23, no. 2, pp. 1184–1199, 2022.
- [3] L. Sun, K. Yang, X. Hu, W. Hu, and K. Wang, “Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images,” *RA-L*, vol. 5, no. 4, pp. 5558–5565, 2020.
- [4] J. Zhang, K. Yang, A. Constantinescu, K. Peng, K. Müller, and R. Stiefelhagen, “Trans4Trans: Efficient transformer for transparent object and semantic scene segmentation in real-world navigation assistance,” *T-ITS*, vol. 23, no. 10, pp. 19 173–19 186, 2022.
- [5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *TPAMI*, vol. 40, no. 4, pp. 834–848, 2018.
- [6] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *CVPR*, 2017.
- [7] J. Fu *et al.*, “Dual attention network for scene segmentation,” in *CVPR*, 2019.
- [8] X. Hu, K. Yang, L. Fei, and K. Wang, “ACNet: Attention based network

- to exploit complementary features for RGBD semantic segmentation," in *ICIP*, 2019.
- [9] X. Chen *et al.*, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *ECCV*, 2020.
- [10] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *IROS*, 2017.
- [11] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, and J. Han, "ABMDRNet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation," in *CVPR*, 2021.
- [12] K. Xiang, K. Yang, and K. Wang, "Polarization-driven semantic segmentation via efficient attention-bridged fusion," *OE*, vol. 29, no. 4, pp. 4802–4820, 2021.
- [13] J. Zhang, K. Yang, and R. Stiefelhagen, "ISSAFE: Improving semantic segmentation in accidents by fusing event-based data," in *IROS*, 2021.
- [14] Z. Zhuang, R. Li, K. Jia, Q. Wang, Y. Li, and M. Tan, "Perception-aware multi-sensor fusion for 3D LiDAR semantic segmentation," in *ICCV*, 2021.
- [15] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, "ShapeConv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation," in *ICCV*, 2021.
- [16] L.-Z. Chen, Z. Lin, Z. Wang, Y.-L. Yang, and M.-M. Cheng, "Spatial information guided convolution for real-time RGBD semantic segmentation," *TIP*, vol. 30, pp. 2313–2324, 2021.
- [17] F. Deng *et al.*, "FEANet: Feature-enhanced attention network for RGB-thermal real-time semantic segmentation," in *IROS*, 2021.
- [18] D. Sun, X. Huang, and K. Yang, "A multimodal vision sensor for autonomous driving," in *SPIE*, 2019.
- [19] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," in *CVPR*, 2022.
- [20] A. Vaswani *et al.*, "Attention is all you need," in *NeurIPS*, 2017.
- [21] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.
- [22] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021.
- [23] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021.
- [24] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *ECCV*, 2012.
- [25] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D," *TPAMI*, vol. 45, no. 3, pp. 3292–3310, 2023.
- [26] D. Gehrig, M. Rüegg, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction," *RA-L*, vol. 6, no. 2, pp. 2822–2829, 2021.
- [27] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," *ICCV*, 2021.
- [28] T. Zhou, W. Wang, E. Konukoglu, and L. Van Gool, "Rethinking semantic segmentation: A prototype view," in *CVPR*, 2022.
- [29] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018.
- [30] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *ICCV*, 2019.
- [31] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *CVPR*, 2021.
- [32] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *ICCV*, 2021.
- [33] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *NeurIPS*, 2021.
- [34] W. Wang *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021.
- [35] Y. Yuan *et al.*, "HRFormer: High-resolution transformer for dense prediction," in *NeurIPS*, 2021.
- [36] Y. Zhang, B. Pang, and C. Lu, "Semantic segmentation by early region proxy," in *CVPR*, 2022.
- [37] F. Lin, Z. Liang, J. He, M. Zheng, S. Tian, and K. Chen, "StructToken : Rethinking semantic segmentation with structural prior," *TCSV7*, 2023.
- [38] Y. Qian, L. Deng, T. Li, C. Wang, and M. Yang, "Gated-residual block for semantic segmentation using RGB-D data," *T-ITS*, vol. 23, no. 8, pp. 11 836–11 844, 2022.
- [39] H. Zhou, L. Qi, H. Huang, X. Yang, Z. Wan, and X. Wen, "CANet: Co-attention network for RGB-D semantic segmentation," *PR*, vol. 124, p. 108468, 2022.
- [40] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *RA-L*, vol. 4, no. 3, pp. 2576–2583, 2019.
- [41] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion," *T-ASE*, vol. 18, no. 3, pp. 1000–1011, 2021.
- [42] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *TIP*, vol. 30, pp. 7790–7802, 2021.
- [43] A. Kalra, V. Taamazyan, S. K. Rao, K. Venkataraman, R. Raskar, and A. Kadambi, "Deep polarization cues for transparent object segmentation," in *CVPR*, 2020.
- [44] J. Zhang, K. Yang, and R. Stiefelhagen, "Exploring event-driven dynamic context for accident scene segmentation," *T-ITS*, vol. 23, no. 3, pp. 2606–2622, 2022.
- [45] W. Wang and U. Neumann, "Depth-aware CNN for RGB-D segmentation," in *ECCV*, 2018.
- [46] Y. Xing, J. Wang, and G. Zeng, "Malleable 2.5D convolution: Learning receptive fields along the depth-axis for RGB-D scene parsing," in *ECCV*, 2020.
- [47] Z. Wu, G. Allibert, C. Stoltz, and C. Demonceaux, "Depth-adapted CNN for RGB-D cameras," in *ACCV*, 2020.
- [48] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in *CVPR*, 2019.
- [49] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "MultiMAE: Multi-modal multi-task masked autoencoders," in *ECCV*, 2022.
- [50] P. Zhang, W. Liu, Y. Lei, and H. Lu, "Hyperfusion-net: Hyper-densely reflective feature fusion for salient object detection," *PR*, vol. 93, pp. 521–533, 2019.
- [51] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "CAVER: Cross-modal view-mixed transformer for bi-modal salient object detection," *TIP*, 2023.
- [52] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *CVPR*, 2017.
- [53] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li, "Efficient attention: Attention with linear complexities," in *WACV*, 2021.
- [54] J. Li, A. Hassani, S. Walton, and H. Shi, "ConvMLP: hierarchical convolutional MLPs for vision," *arXiv preprint arXiv:2109.04454*, 2021.
- [55] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *ECCV*, 2014.
- [56] R. Yan, K. Yang, and K. Wang, "NLFFNet: Non-local fusion towards generalized multimodal semantic segmentation across RGB-depth, polarization, and thermal images," in *ROBIO*, 2021.
- [57] I. Alonso and A. C. Murillo, "EV-SegNet: Semantic segmentation for event-based cameras," in *CVPRW*, 2019.
- [58] E. Mohammadbagher, N. P. Bhatt, E. Hashemi, B. Fidan, and A. Khajepour, "Real-time pedestrian localization and state estimation using moving horizon estimation," in *ITSC*, 2020.
- [59] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *CVPR*, 2015.
- [60] G. Zhang, J.-H. Xue, P. Xie, S. Yang, and G. Wang, "Non-local aggregation for RGB-D semantic segmentation," *SPL*, vol. 28, pp. 658–662, 2021.
- [61] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2D-3D-semantic data for indoor scene understanding," *arXiv preprint arXiv:1702.01105*, 2017.
- [62] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *CVPR*, 2017.
- [63] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.
- [64] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *CoRL*, 2017.
- [65] Z. Sun, N. Messikommer, D. Gehrig, and D. Scaramuzza, "ESS: Learning event-based semantic segmentation from still images," in *ECCV*, 2022.
- [66] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.
- [67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
- [68] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3D graph neural networks for RGBD semantic segmentation," in *ICCV*, 2017.

- [69] S. Kong and C. C. Fowlkes, “Recurrent scene parsing with perspective understanding in the loop,” in *CVPR*, 2018.
- [70] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, “Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation,” in *CVPR*, 2017.
- [71] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, and H. Huang, “Cascaded feature network for semantic segmentation of RGB-D images,” in *ICCV*, 2017.
- [72] S.-J. Park, K.-S. Hong, and S. Lee, “RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation,” in *ICCV*, 2017.
- [73] F. Fooladgar and S. Kasaei, “Multi-modal attention-based fusion model for semantic segmentation of RGB-depth images,” *arXiv preprint arXiv:1912.11691*, 2019.
- [74] Y. Yue, W. Zhou, J. Lei, and L. Yu, “Two-stage cascaded decoder for semantic segmentation of RGB-D images,” *SPL*, vol. 28, pp. 1115–1119, 2021.
- [75] A. Valada, R. Mohan, and W. Burgard, “Self-supervised model adaptation for multimodal semantic segmentation,” *IJCV*, vol. 128, no. 5, pp. 1239–1285, 2019.
- [76] A. Dai and M. Nießner, “3DMV: Joint 3D-multi-view prediction for 3D semantic scene segmentation,” in *ECCV*, 2018.
- [77] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, “FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture,” in *ACCV*, 2016.
- [78] W. Shi *et al.*, “Multilevel cross-aware RGBD indoor semantic segmentation for bionic binocular robot,” *T-MRB*, vol. 2, no. 3, pp. 382–390, 2020.
- [79] W. Shi *et al.*, “RGB-D semantic segmentation and label-oriented voxelgrid fusion for accurate 3D semantic mapping,” *TCSVT*, vol. 32, no. 1, pp. 183–197, 2022.
- [80] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic, “In defense of pre-trained ImageNet architectures for real-time semantic segmentation of road-driving images,” in *CVPR*, 2019.
- [81] D. Seichter, M. Köhler, B. Lewandowski, T. Wengfeld, and H.-M. Gross, “Efficient RGB-D semantic segmentation for indoor scene analysis,” in *ICRA*, 2021.
- [82] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, “Gated-SCNN: Gated shape CNNs for semantic segmentation,” in *ICCV*, 2019.
- [83] F. Zhang *et al.*, “ACFNet: Attentional class feature network for semantic segmentation,” in *ICCV*, 2019.
- [84] D. Xu, W. Ouyang, X. Wang, and N. Sebe, “PAD-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing,” in *CVPR*, 2018.
- [85] Y. Wang, F. Sun, M. Lu, and A. Yao, “Learning deep multimodal feature representation with asymmetric multi-layer fusion,” in *MM*, 2020.
- [86] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016.
- [87] X. Zhang, S. Zhang, Z. Cui, Z. Li, J. Xie, and J. Yang, “Tube-embedded transformer for pixel prediction,” *TMM*, vol. 25, pp. 2503–2514, 2023.
- [88] W. Hu, H. Zhao, L. Jiang, J. Jia, and T.-T. Wong, “Bidirectional projection network for cross dimension scene understanding,” in *CVPR*, 2021.
- [89] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, “ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation,” *T-ITS*, vol. 19, no. 1, pp. 263–272, 2018.
- [90] J. Wang *et al.*, “Deep high-resolution representation learning for visual recognition,” *TPAMI*, vol. 43, no. 10, pp. 3349–3364, 2021.
- [91] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, “PST900: RGB-thermal calibration, dataset and segmentation network,” in *ICRA*, 2020.
- [92] J. Xu, K. Lu, and H. Wang, “Attention fusion network for multi-spectral semantic segmentation,” *PRL*, vol. 146, pp. 179–184, 2021.
- [93] Y. Cai, W. Zhou, L. Zhang, L. Yu, and T. Luo, “DHFNet: Dual-decoding hierarchical fusion network for RGB-thermal semantic segmentation,” *The Visual Computer*, pp. 1–11, 2023.
- [94] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, “Full-resolution residual networks for semantic segmentation in street scenes,” in *CVPR*, 2017.
- [95] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “Learning a discriminative feature network for semantic segmentation,” in *CVPR*, 2018.
- [96] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “BiSeNet: Bilateral segmentation network for real-time semantic segmentation,” in *ECCV*, 2018.
- [97] R. P. K. Poudel, S. Liwicki, and R. Cipolla, “Fast-SCNN: Fast semantic segmentation network,” in *BMVC*, 2019.
- [98] T. Wu, S. Tang, R. Zhang, and Y. Zhang, “CGNet: A light-weight context guided network for semantic segmentation,” *TIP*, vol. 30, pp. 1169–1179, 2021.
- [99] J. Zhang, K. Yang, A. Constantinescu, K. Peng, K. Müller, and R. Stiefelhagen, “Trans4Trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world,” in *ICCVW*, 2021.
- [100] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *ECCV*, 2018.
- [101] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, “Unified perceptual parsing for scene understanding,” in *ECCV*, 2018.
- [102] T. Brodernmann, C. Sakaridis, D. Dai, and L. Van Gool, “HRFuser: A multi-resolution sensor fusion architecture for 2D object detection,” in *ITSC*, 2023.
- [103] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, “Multi-modal token fusion for vision transformers,” in *CVPR*, 2022.
- [104] A. Prakash, K. Chitta, and A. Geiger, “Multi-modal fusion transformer for end-to-end autonomous driving,” in *CVPR*, 2021.

## APPENDIX A MORE IMPLEMENTATION DETAILS

We implement our experiments with PyTorch. We employ a poly learning rate schedule with a factor of 0.9 and an initial learning rate of  $6e^{-5}$ . The number of warm-up epochs is 10. We now describe implementation details for different datasets.

**NYU Depth V2 dataset.** We train our model with the MiT-B2 backbone on four 2080Ti GPUs, models with MiT-B4 and MiT-B5 backbones on three 3090 GPUs. The number of training epochs is set as 500. We take the whole image with the size  $640 \times 480$  for training and inference. We use a batch size of 8 for the MiT-B2 backbone and 6 for MiT-B4 and -B5. **SUN-RGBD dataset.** The models are trained with a batch size of 4 per GPU. During training, the images are randomly cropped to  $480 \times 480$ . The model based on MiT-B2 is trained on two V100 GPUs for 200 epochs. The models based on MiT-B4 and MiT-B5 are trained on eight V100 GPUs, 250 epochs for MiT-B4 and 300 epochs for MiT-B5.

**Stanford2D3D dataset.** The model is trained on four 2080Ti GPUs. The number of training epochs here is set as 32. We resize the input images to  $480 \times 480$ . We use a batch size of 12 for the MiT-B2 backbone and 8 for MiT-B4.

**ScanNetV2 dataset.** The model is trained on four 2080Ti GPUs. The number of training epochs here is set as 100. We resize the input RGB images to  $640 \times 480$ . We use a batch size of 12 for the MiT-B2 backbone.

**Cityscapes dataset.** The model is trained on eight A100 GPUs for 500 epochs. The batch size is set as 8. The images are randomly cropped into  $1024 \times 1024$  for training and inference is performed on the full resolution with a sliding window of  $512 \times 512$ . The embedding dimension of the MiT-B4 backbone and MLP-decoder is set as 768.

**RGB-T MFNet dataset.** The model is trained on four 2080Ti GPUs. We use the original image size of  $640 \times 480$  for training and inference. The batch size is set to 8 for the MiT-B2 backbone and we train for 500 epochs. Consistent with the batch size of 8, the model based on MiT-B4 is trained on four A100 GPUs, which requires a larger memory.

**RGB-P ZJU dataset.** The model is trained on four 2080Ti GPUs. We resize the image from  $1224 \times 1024$  to  $612 \times 512$ . The number of training epochs is set as 400. We use a batch size of 8 for the MiT-B2 backbone and 4 for MiT-B4. In practice, we calculate the image encoding pixel-wise *AoLP* information by mapping the values of  $\arctan(S_1/S_2)$  to the range of  $[0, 255]$ .

**RGB-E EventScape dataset.** The proposed model is trained with a batch size of 4 and the original resolution of  $512 \times 256$  on a single 1080Ti GPU. The number of training epochs is set as 100. The embedding dimension of the MiT-B4 backbone and MLP-decoder is set as 768.

**RGB-L KITTI-360 dataset.** The model is trained with a batch size of 2 and the original resolution of  $1408 \times 376$ . The number of training epochs is set as 40.

## APPENDIX B MORE QUALITATIVE ANALYSIS

**Segmentation results on the Cityscapes dataset.** We further view the outdoor RGB-D semantic segmentation results on

the Cityscapes dataset based on the backbone of SegFormer-B4. We show the results of the RGB-only baseline and our RGB-X approach, in particular, the difference maps w.r.t. the segmentation ground truth. As displayed in Fig. A.1, in spite of the noisy depth measurements, our CMX still benefits from the HHA-encoded image, thanks to the ability to rectify and fuse cross-modal complementary features. Our approach has higher pixel accuracy scores on a wide variety of driving scene elements such as *fence*, and *sidewalk* in the positive group (in green boxes). However, the shadows and weak illumination conditions are still challenging for both models and make the depth cues less effective. For example, depth information in the regions of *sidewalk* in the negative group (in red boxes), may be less informative for fusion.

**Failure case analysis.** In Fig. A.2, we show a set of failure cases in different sensing modality combination scenarios. The first row shows that for the RGB-D semantic segmentation in a highly composite indoor scene with extremely densely arranged objects, the parsing results are still less visually satisfactory. In the second row of a nighttime scene, the *guardrails* are misclassified by the RGB-X method as *color cone*, despite our model delivering more complete and consistent segmentation than the RGB-only model and having better segmentation of *person* with thermal properties. This illustrates that at night, the perception of some remote objects is still challenging in RGB-T semantic segmentation and it should be noted for safety-critical applications like automated driving. In the third row, the RGB-P model might be misguided by the polarized background area in an occluded situation and yields less accurate parsing results, indicating that polarization, as a strong prior for segmentation of specular surfaces like *glass* and *car* regions, should be carefully leveraged in unconstrained scenes with a lot of occlusions. In the fourth row, the *fences* are partially detected as *vehicles* in the RGB-E segmentation result, but our model still yields more correctly identified pixels than the RGB-only model by harvesting complementary cues from event data. In the last row, the over-exposed *sidewalk* region is still a challenge for segmentation. Nonetheless, our RGB-LiDAR CMX predicts a much better mask on the *fence* region, where the spatial information given by LiDAR data is more accurate.

**Feature analysis.** To understand the key module for feature rectification, we visualize the input- and rectified features of CM-FRM in layer 1, and their difference map, as shown in Fig. B.1. It can be seen that the feature maps are enhanced in both streams after the cross-modal calibration. The RGB stream delivers texture information to the supplement modality, while the supplement modality further improves the boundary and emphasizes complementary discontinuities of RGB features. In the RGB-D segmentation scenario, the RGB-feature difference map shows that the ground area is better spotlighted, thanks to the HHA image encoding depth information, which provides geometric cues such as height above ground, beneficial for higher-level semantic prediction of ground-related classes. In the RGB-T nighttime scene parsing cases, the pedestrians are hard to be seen in the RGB images. But the RGB-feature difference map clearly highlights the pedestrians thanks to the supplementary thermal modality

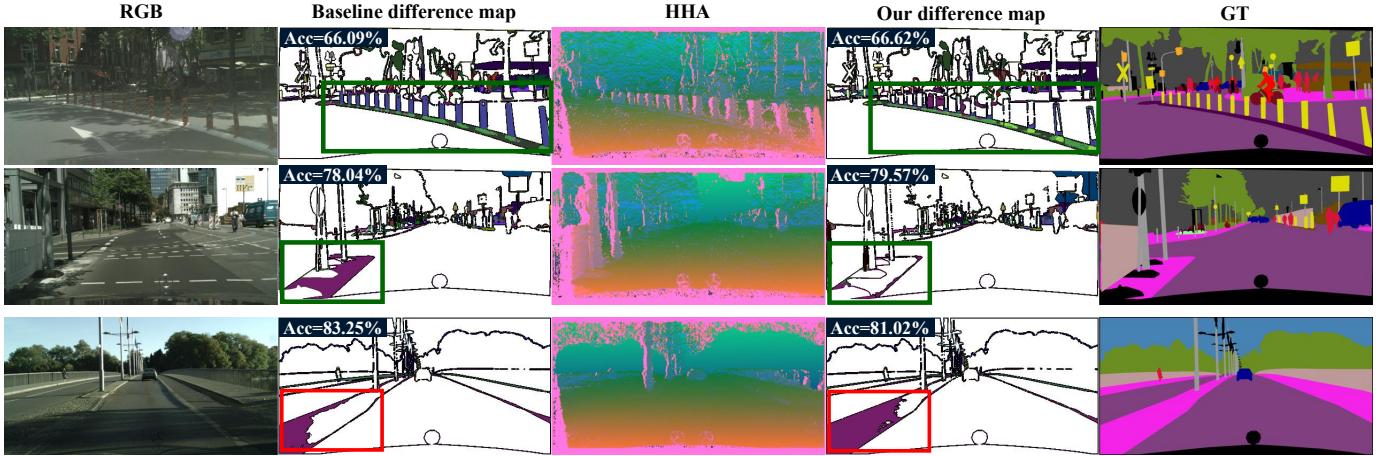


Fig. A.1: Visualization of semantic segmentation results for the RGB-only baseline and our RGB-X approach, both of which are based on SegFormer-B4. “Acc” is short for pixel accuracy of the segmentation result. From left to right: RGB image, baseline difference map *w.r.t.* the ground truth, HHA image encoding depth information, our difference map, and ground truth.

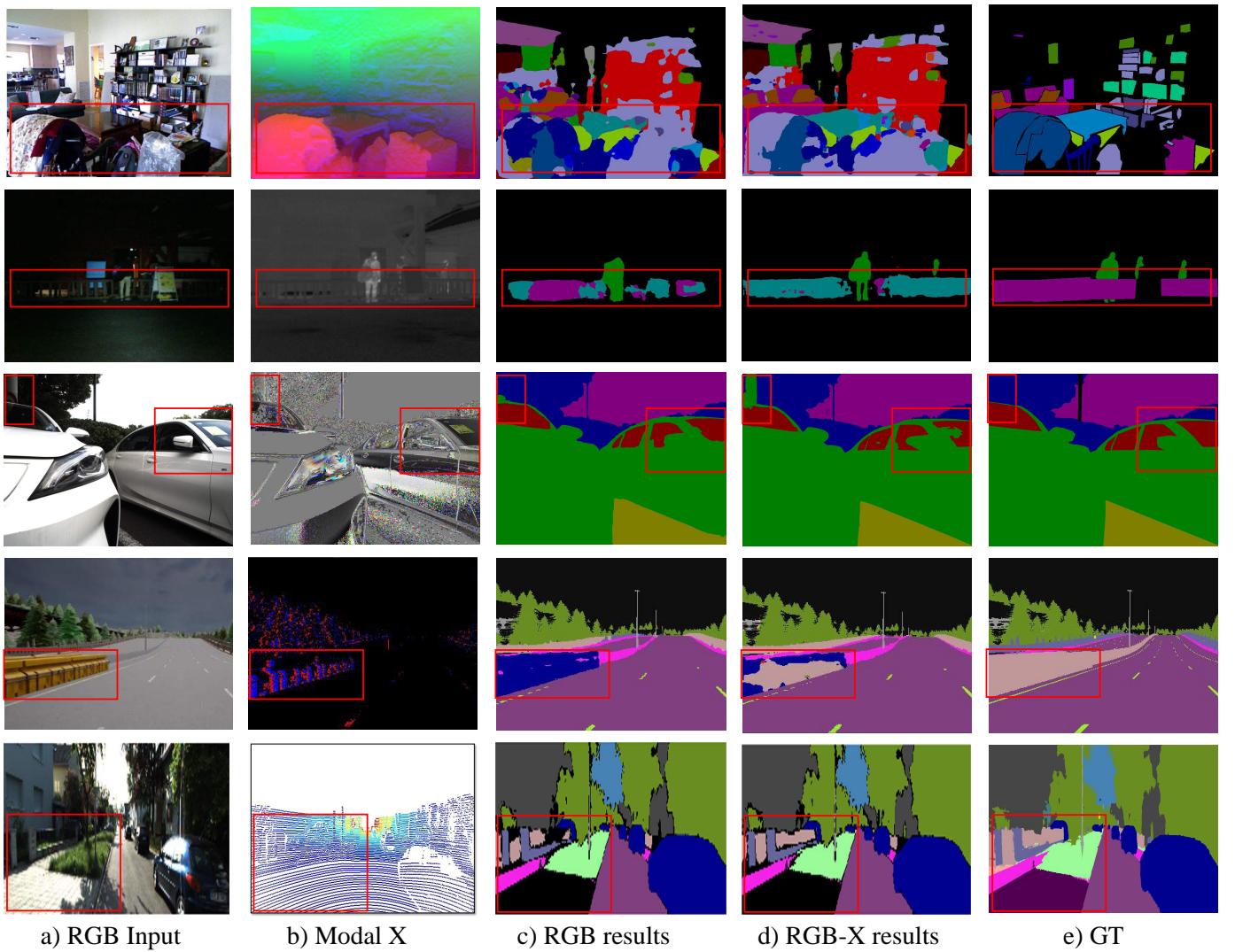


Fig. A.2: Visualization of failure cases. We use SegFormer-B2 for RGB segmentation and the proposed approach with the same backbone Mit-B2 and MLP-Decoder for RGB-X segmentation. From top to bottom: RGB-Depth, RGB-Thermal, RGB-Polarization (AoLP), and RGB-Event semantic segmentation.

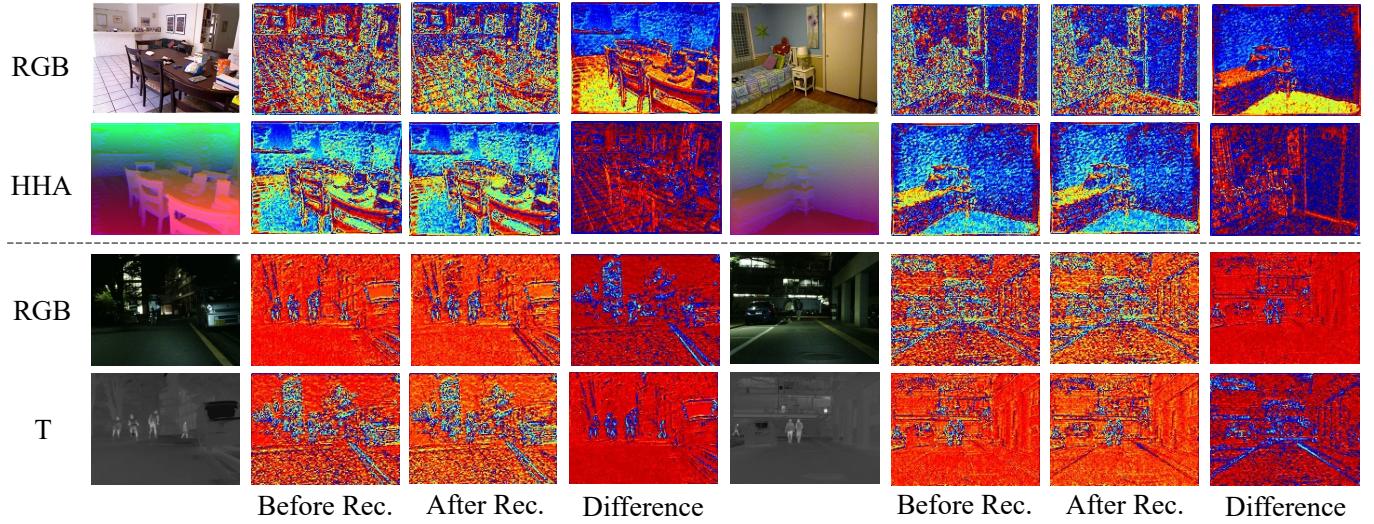


Fig. B.1: Visualization of the feature extracted in layer 1 and the rectified feature, and their difference map.

with infrared imaging. These indicate that the complementary features have been infused into the RGB stream. The RGB features have been rectified to better focus on informative ones and capture such complementary discontinuities towards accurate semantic understanding.