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

Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen, Member, IEEE

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 state-of-the-art. The source code of CMX is publicly available at https://github.com/huaaaliu/RGBX_Semantic_Segmentation.

Index Terms—Semantic segmentation, scene parsing, cross-modal fusion, vision transformers, scene understanding.

I. INTRODUCTION

SCENE 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 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 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]...
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.

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.

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.

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 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 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
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 non-parametric 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.

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.

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

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.
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.

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

\[
\text{W}^C_{\text{RGB}}, \text{W}^C_{\text{X}} = \mathcal{F}_{\text{split}}(\sigma(\mathcal{F}_{\text{mlp}}(\text{Y}))).
\]

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

\[
\text{RGB}_{\text{rec}} = \text{W}^C_{\text{RGB}} \odot \text{X}_n, \quad \text{X}_{\text{rec}} = \text{W}^C_{\text{RGB}} \odot \text{RGB}_n,
\]

where \( \odot \) denotes channel-wise multiplication.

2) 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 \( \text{RGB}_n \) and \( \text{X}_n \) will be concatenated and embedded into two spatial weight maps: \( \text{W}^S_{\text{RGB}} \in \mathbb{R}^{H \times W} \) and \( \text{W}^S_{\text{X}} \in \mathbb{R}^{H \times W} \). The embedding operation has two 1x1 convolution layers assembled with a RELU function. Afterward, a sigmoid function is applied to obtain the embedded feature map \( \text{F} \in \mathbb{R}^{H \times W \times 2} \), which is further split into two weight maps. The process to obtain the spatial weight maps is formulated as:

\[
\text{F} = \text{Conv}_{1 \times 1}(\text{RELU}(\text{Conv}_{1 \times 1}((\text{RGB}_n \parallel \text{X}_n))))
\]

\[
\text{W}^S_{\text{RGB}}, \text{W}^S_{\text{X}} = \mathcal{F}_{\text{split}}(\sigma(\text{F})).
\]

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

\[
\text{RGB}_{\text{rec}} = \text{W}^S_{\text{RGB}} \ast \text{X}_n, \quad \text{X}_{\text{rec}} = \text{W}^S_{\text{RGB}} \ast \text{RGB}_n,
\]

where \( \ast \) denotes spatial-wise multiplication.

The whole rectified feature for both modalities \( \text{RGB}_{\text{out}} \) and \( \text{X}_{\text{out}} \) is organized as:

\[
\text{RGB}_{\text{out}} = \text{RGB}_n + \lambda_C \text{RGB}_{\text{rec}} + \lambda_S \text{RGB}_{\text{rec}}, \quad \text{X}_{\text{out}} = \text{X}_n + \lambda_C \text{X}_{\text{rec}} + \lambda_S \text{X}_{\text{rec}},
\]

\( \lambda_C \) and \( \lambda_S \) are two hyperparameters. We set them both as 0.5 as default and will ablative in Sec. V-F. \( \text{RGB}_{\text{out}} \) and \( \text{X}_{\text{out}} \) 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.

1) 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 \( \mathbb{R}^{N \times C} \), which we call residual vector \( \text{X}_{\text{res}} \) and interactive vector \( \text{X}_{\text{inter}} \). 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 (Q), Key (K), and Value (V). The global attention map is calculated via a matrix multiplication \( QK^T \), which has a size of \( \mathbb{R}^{N \times N} \) and causes a high memory occupation. In contrast, [53] uses a global context vector \( \text{G} = \text{K}^T \text{V} \) with a size \( \mathbb{R}^{\text{C}_{\text{head}} \times \text{C}_{\text{head}}} \) and the attention result is calculated by \( \text{QG} \). 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 K and V for each head, and both sizes of them are \( \mathbb{R}^{\text{N} \times \text{C}_{\text{head}}} \). 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:

\[
\text{G}_{\text{RGB}} = \text{K}_{\text{RGB}} \text{V}_{\text{RGB}}, \quad \text{G}_{\text{X}} = \text{K}_{\text{X}} \text{V}_{\text{X}},
\]

\[
\text{U}_{\text{RGB}} = \text{X}^\text{inter}_{\text{RGB}} \text{SoftMax}(\text{G}_{\text{X}}), \quad \text{U}_{\text{X}} = \text{X}^\text{inter}_{\text{X}} \text{SoftMax}(\text{G}_{\text{RGB}}).
\]

Note that \( \text{G} \) denotes the global context vector, while \( \text{U} \) 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 \( \text{U} \) and the residual vector \( \text{X}_{\text{res}} \) are concatenated. Finally, we apply a second linear embedding and resize the feature to \( \mathbb{R}^{H \times W \times C} \).
2) Fusion Stage: In the second stage of FFM, precisely the fusion stage, we use a single 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

1) 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.

2) 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.

3) 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 $I_{\text{angle}}$ 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:

$$S_0 = I_0^\circ \quad S_1 = I_{45}^\circ \quad S_2 = I_{90}^\circ \quad S_3 = I_{135}^\circ.$$ 

Then, DoLP and AoLP are formally computed as:

$$\text{DoLP} = \frac{\sqrt{S_1^2 + S_2^2}}{S_0},$$

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

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.

4) 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.

5) 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:

$$f_x = H/(2 \times \tan(FoV \times \pi/360)),$$

$$f_y = W/(2 \times \tan(FoV \times \pi/360)). \quad (12)$$

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

$$\begin{bmatrix}
    u \\
    v \\
    1
\end{bmatrix} =
\begin{bmatrix}
    f_x & 0 & u_0 & 0 \\
    0 & f_y & v_0 & 0 \\
    0 & 0 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
    X \\
    Y \\
    Z \\
    1
\end{bmatrix}, \quad (13)$$
where \((X, Y, Z)\) is the LiDAR point, \((u, v)\) is the 2D image pixel, and the rotation \((R)\) and the translation \((t)\) matrices are given by KITTI-360 dataset [25].

IV. EXPERIMENTAL SETTINGS

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\(^1\) 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, 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 \(6 \times 10^{-5}\) 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.

---

\(^1\)https://paperswithcode.com/sota/semantic-segmentation-on-eventscape

| Dataset          | Image   | Event | Train/Val | Label   | Resolution | Class |
|------------------|---------|-------|-----------|---------|------------|-------|
| DDDV17 [17]      | Gray-scale | SUN   | 500/150   | pseudo  | 346 x 260  | 6     |
| DDBC-Semantic [65] | Gray-scale | SUN   | 8082/2809 | pseudo  | 640 x 440  | 11    |
| EventScape [26]  | RGB     | SUN   | 12329/2249| synthetic| 512 x 256  | 12    |

---
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 [69] | 43.1     | -       |
| Kong et al. [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    |
| SGNNet [16] | 51.1     | 76.8    |
| ShapeConv [15] | 51.3     | 76.4    |
| NaNNet [60] | 52.3     | 77.9    |
| SA-Gate [9] | 52.4     | 77.9    |
| CMX (MiT-B2) | 54.1     | 78.7    |
| CMX (MiT-B2*) | 54.4    | 79.9    |
| CMX (MiT-B4) | 56.0     | 79.6    |
| CMX (MiT-B4*) | 56.3    | 79.9    |
| CMX (MiT-B5) | 56.8     | 79.9    |
| CMX (MiT-B5*) | 56.9    | 80.1    |

(b) Results on Stanford2D3D [61].

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

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

| Method      | mIoU (%) | Acc (%) |
|------------|----------|---------|
| 3DGNN [69] | 49.9     | -       |
| RDF-152 [72] | 47.7    | 81.5    |
| CFN [71]   | 48.1     | -       |
| ACNet [8]  | 48.1     | -       |
| TCD [74]   | 49.5     | 83.1    |
| SGNNet [16] | 46.6     | 82.0    |
| SA-Gate [9] | 49.4     | 82.5    |
| NaNNet [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) | 52.4     | 83.8    |

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

| Method      | mIoU (%) | Acc (%) |
|------------|----------|---------|
| PSPNet [4] | 47.5     | -       |
| AdapNet++ [75] | 50.3   | -       |
| 3DMV (2d-proj) [76] | RGB-D | 49.8   |
| PaseNet [77] | RGB-D | 53.5   |
| SSMA [75]  | RGB-D    | 57.7    |
| GRNM [38]  | RGB-D    | 59.2    |
| MCA-Net [78] | RGB-D | 59.5    |
| DMMF [79]  | RGB-D    | 59.7    |
| CMX (MiT-B2) | RGB-D   | 61.3    |

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

| Method      | Modal | Backbone | mIoU (%) |
|------------|-------|----------|----------|
| SwifNet [80] | RGB   | ResNet-18 | 70.4    |
| ESA-Net [81] | RGB   | ResNet-50 | 87.2    |
| GSCNN [82]  | RGB   | WideResNet-38 | 80.8    |
| CCNet [80]  | RGB   | ResNet-101 | 81.3    |
| DANE [7]    | RGB   | ResNet-101 | 81.5    |
| ACPNet [63] | RGB   | ResNet-101 | 81.5    |
| SegFormer [33] | RGB   | MiT-B2   | 81.0    |
| SegFormer [33] | RGB   | MiT-B4   | 82.3    |

A. Results on RGB-D Depth Datasets

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

1) 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 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.

2) 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 61.9%. The results demonstrate the effectiveness and learning capacity of our approach on such a large RGB-D dataset.

3) 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., SGNNet [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].

4) 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.

5) 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

1) 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-D 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 var-
TABLE III

| 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        | 30.8 | 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 |
| PSNNet [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 |

TABLE IV

| Method          | Modal | Daytime mIoU (%) | Nighttime mIoU (%) |
|-----------------|-------|------------------|--------------------|
| FRN [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               |
| PSENet [41]     | RGB-T | 41.0             | 43.9               |
| RTFNet [40]     | RGB-T | 45.8             | 45.8               |
| PSESeg [41]     | RGB-T | 47.8             | 54.6               |
| GMNet [42]      | RGB-T | 49.0             | 57.1               |
| CMX (MiT-B2)    | RGB-T | 51.3             | 57.8               |
| CMX (MiT-B4)    | RGB-T | 52.5             | 59.4               |

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

2) 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

1) 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.

2) 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

1) 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],
TABLE V
PER-CLASS RESULTS ON ZJU-RGB-P DATASET [12] FOR RGB-POLARIZATION SEGMENTATION

| Method          | Modal | Building (%) | Glass (%) | Car (%) | Road (%) | Vegetation (%) | Sky (%) | Pedestrian (%) | Bicycle (%) | mIoU (%) |
|-----------------|-------|--------------|-----------|---------|----------|----------------|---------|----------------|-------------|----------|
| SwiftNet [89]   | 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    | 98.6     | 96.2           | 89.6    | 82.9           | 89.3        | 89.6     |
| NLNet [56]      | RGB-P | 85.4         | 77.1      | 93.5    | 97.7     | 93.2           | 85.9    | 56.9           | 85.5        | 84.4     |
| EAPNet [12]     | RGB-P | 87.0         | 79.3      | 93.6    | 97.4     | 95.3           | 87.1    | 60.4           | 85.8        | 85.7     |
| CMX (SegFormer-B2) | RGB-AoLP (Monochromatic) | 91.9      | 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      | 88.8      | 96.3    | 98.3     | 96.7           | 89.1    | 86.3           | 92.3        | 92.4     |
| CMX (SegFormer-B4) | RGB-AoLP (Trichromatic) | 91.6      | 88.8      | 96.3    | 98.3     | 96.8           | 89.7    | 86.2           | 92.8        | 92.6     |

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           | 89.4    | 86.1           | 91.8        | 92.2     |
| CMX (SegFormer-B4) | RGB-DoLP (Monochromatic) | 91.8      | 88.6      | 96.3    | 98.3     | 96.7           | 89.4    | 86.0           | 92.1        | 92.4     |
| CMX (SegFormer-B4) | RGB-DoLP (Trichromatic) | 91.6      | 88.6      | 96.3    | 98.3     | 96.7           | 89.5    | 86.4           | 92.2        | 92.5     |

TABLE VI
RESULTS FOR RGB-EVENT SEGMENTATION

| Method          | Modal | Backbone | mIoU (%) | Pixel Acc. (%) |
|-----------------|-------|----------|----------|----------------|
| SwiftNet [89]   | RGB   | ResNet-18 | 36.67    | 83.46         |
| Fast-SCNN [97]  | RGB   | Fast-SCNN | 44.27    | 87.10         |
| GCNet [98]      | RGB   | M32N1     | 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    | 64.28    | 92.60         |

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

[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%).

2) 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%, +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.

3) 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], and [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...
hensively compare the RGB-only baseline [33] and our segmentation performance is evaluated on NYU Depth V2. We take MiT-B2 as the backbone with the MLP modality here. We take MiT-B2 as the backbone with the MLP feature in modal fusion and the effectiveness of our proposed feature in modal fusion and the effectiveness of our proposed.

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.

1) 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.

2) 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.

3) 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.

4) 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
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) | 1.6 (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          |
| ✓           | ✓    | 54.1     | 78.7          |

TABLE X
ABLATION WITH CM-FRM/FFM VARIANTS ON NYU DEPTH V2 TEST SET

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

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 | 51.1    | 75.7          |
| RGB + HHA     | 52.0     | 77.0          |

H. Qualitative Analysis

1) 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 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.

2) 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.

3) 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 a truck. Furthermore, our

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 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.

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

| Method                 | #Params (M) | FLOPs (G) | mIoU (%) |
|------------------------|-------------|-----------|----------|
| SA-Gate [9] (ResNet50) | 63.4        | 204.9     | 50.4     |
| CMX (SegFormer-B2)     | 66.6        | 67.6      | 54.1     |
| CMX (SegFormer-B4)     | 139.9       | 134.3     | 56.0     |
| CMX (SegFormer-B5)     | 181.1       | 167.8     | 56.8     |
CM-FRM 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 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,” IEEE Trans. Intell. Transp. Syst., vol. 21, no. 5, pp. 1951–1963, May 2020.
[2] K. Yang, X. Hu, Y. Fang, K. Wang, and R. Stiefelhagen, “Omnisupervised omnidirectional semantic segmentation,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 2, pp. 1184–1199, Feb. 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,” IEEE Robot. Autom. Lett., vol. 5, no. 4, pp. 5558–5565, Oct. 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,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 10, pp. 19173–19186, Oct. 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,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2018.
Jiaming Zhang received the B.Sc. degree in computer science and software engineering from Shenzhen University (SZU) in 2015, and the M.Sc. degree in computer science from the Karlsruhe Institute of Technology (KIT) in 2020. He is currently a Research Assistant and a Ph.D. candidate at the Computer Vision for Human-Computer Interaction (CV:HCI) Lab at KIT. In 2023, he joined the Torr Vision Group (TVG) at the University of Oxford as a visiting Ph.D. student. His research fields include scene understanding, visual relocalization, and their applications in intelligent vehicles and assistive systems for people with visual impairments.

Huayao Liu received the B.Sc degree in mechatronics from Tongji University in 2018, and the master’s degree in mechatronics and information technology from the Karlsruhe Institute of Technology (KIT) in 2022. He is writing his master thesis at the Computer Vision for Human-Computer Interaction (CV:HCI) Lab at the Institute for Anthropomatics and Robotics (IAR) at KIT. He is currently an Algorithm Researcher at NIO Ltd. His research interests include computer vision, robotics, digital cabin technologies, and multi-modal Human-Computer Interaction.

Kailun Yang received the dual B.S. degrees in measurement technology and instrument from the Beijing Institute of Technology (BIT) and Economics from Peking University (PKU) in 2014, and the Ph.D. degree in information sensing and instrumentation from the State Key Laboratory of Extreme Photonics and Instrumentation, Zhejiang University (ZJU) in 2019. He performed a Ph.D. internship with the Robotics and eSafety (RobeSafe) research group at the University of Alcalá (UAH) from 2017 to 2018. He was a Post-Doctoral Researcher with the Computer Vision for Human-Computer Interaction (CV:HCI) lab at Karlsruhe Institute of Technology (KIT) from 2019 to 2023. He is currently an Associate Professor with the School of Robotics and the National Engineering Research Center of Robot Visual Perception and Control Technology, Hunan University (HNU).

Xinlin Hu received his B.S. and M.S. degrees from the College of Optical Science and Engineering, Zhejiang University in 2017 and 2020. He has internship experience in central media research institute of Huawei and ArcSoft, mainly engaged in knowledge distillation and portrait segmentation. He is currently a researcher in computer vision with ByteDance Inc. His research interests include optical detection, 3D vision, semantic segmentation, knowledge distillation, and indoor navigation.

Ruiping Liu received the B.Sc. degree in mechatronics from Tongji University in 2018, and the M.Sc. degree in mechatronics and information technology from the Karlsruhe Institute of Technology (KIT) in 2022. She is currently a Research Assistant and a Ph.D. candidate at the Computer Vision for Human-Computer Interaction (CV:HCI) Lab at KIT. She completed internships at Shanghai Shen-Zhou Vehicle Energy and Environmental Protection Co. and Bosch Rexroth. Her research interests include computer vision, deep learning, semantic segmentation, and real-time applications.

Rainer Stiefelhagen (Member, IEEE) received the Diploma (Dipl.-Inform) and Doctoral degree (Dr.-Ing.) from the Universität Karlsruhe (TH) in 1996 and 2002, respectively. He is currently a Full Professor for "Information Technology Systems for Visually Impaired students" at the Karlsruhe Institute of Technology (KIT), where he directs the Computer Vision for Human-Computer Interaction Lab at the Institute for Anthropomatics and Robotics as well as Center for Digital Accessibility and Assistive Technology. His research interests include computer vision methods for visual perception of humans and their activities, in order to facilitate perceptive multimodal interfaces, humanoid robots, smart environments, multimedia analysis, and assistive technology for persons with visual impairments.