Centering the Value of Every Modality: Towards Efficient and Resilient Modality-agnostic Semantic Segmentation

Xu Zheng[1], Yuanhuiyi Lyu[1], Jiazhou Zhou[1], and Lin Wang[1,2]*

[1] Hong Kong University of Science and Technology, Guangzhou, China
zhengxu128@gmail.com, {yuanhuiyilv, jiazhouzhou}@hkust-gz.edu.cn
[2] Hong Kong University of Science and Technology, Hong Kong, China
linwang@ust.hk
https://vlislab22.github.io/MAGIC/

Abstract. Fusing an arbitrary number of modalities is vital for achieving robust multi-modal fusion of semantic segmentation yet remains less explored to date. Recent endeavors regard RGB modality as the center and the others as the auxiliary, yielding an asymmetric architecture with two branches. However, the RGB modality may struggle in certain circumstances, *e.g.*, nighttime, while others, *e.g.*, event data, own their merits; thus, it is imperative for the fusion model to discern robust and fragile modalities, and incorporate the most robust and fragile ones to learn a resilient multi-modal framework. To this end, we propose a novel method, named **MAGIC**, that can be flexibly paired with various backbones, ranging from compact to high-performance models. Our method comprises two key plug-and-play modules. Firstly, we introduce a multi-modal aggregation module to efficiently process features from multi-modal batches and extract complementary scene information. On top, a unified arbitrary-modal selection module is proposed to utilize the aggregated features as the benchmark to rank the multi-modal features based on the similarity scores. This way, our method can eliminate the dependence on RGB modality and better overcome sensor failures while ensuring the segmentation performance. Under the commonly considered multi-modal setting, our method achieves state-of-the-art performance while reducing the model parameters by **60%**. Moreover, our method is superior in the novel modality-agnostic setting, where it outperforms prior arts by a large margin of +**19.41**% mIoU.

Keywords: Semantic Segmentation, Multi-modal Learning, Modality-agnostic Segmentation

1 Introduction

Nature has elucidated that miscellaneous sense and processing capabilities of visual information are vital for the understanding of the sophisticated environ-

* Corresponding author

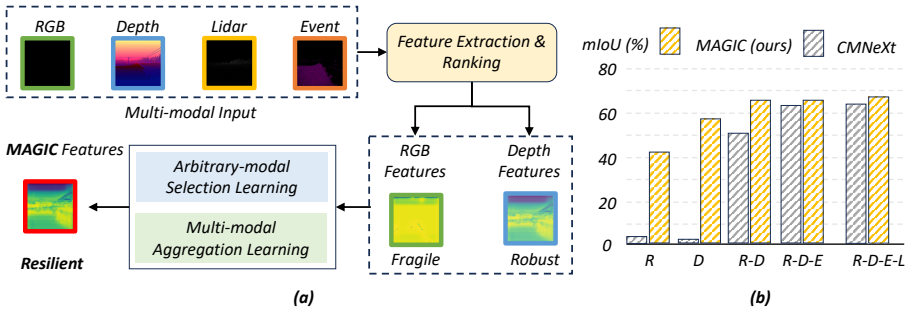

Fig. 1: (a): Robust and fragile features extracted from multi-modal data and the resilient one learned by MAGIC. (b): Modality-agnostic segmentation with arbitrary inputs using {**R**GB, **D**epth, **E**vent, **L**iDAR} on DELIVER.

ment [17,58]. As such, advanced robots or self-driving vehicles need multi-sensor systems, which encompass diverse sensors, such as RGB, LiDAR, and event cameras, to achieve reliable scene understanding, including semantic segmentation [1, 25, 29, 49, 53, 59, 64, 71, 92]. The intuition is that each sensor delivers its distinct advantages [67, 95].

Early endeavors predominantly focus on developing on bespoke fusion methods for specific sensors (*e.g.*, RGB-depth [56], RGB-Lidar [33], RGB-event [88], RGB-thermal [27]). Unfortunately, these methods lack flexibility and versatility when incorporating additional sensors. As a result, fusing an arbitrary number of modalities is trendily desirable for achieving more robust multi-modal fusion in semantic segmentation. Despite its importance, this research direction remains relatively unexplored. Only recently, a few works have been proposed, which regard the RGB modality as the primary while others as the auxiliary [86,87,107]. Naturally, a unified RGB-X pipeline is developed, yielding a distributed joint branch or an asymmetric architecture featuring two branches. In particular, CMNeXt [87] introduces a self-query hub to extract effective information from any auxiliary modalities for subsequent fusion with the RGB modality.

Motivation: Nevertheless, the RGB modality may struggle in certain circumstances, as demonstrated by the visualized feature in nighttime condition in Fig. 1 (a). By contrast, alternative sensors offer distinct benefits that bolster scene understanding in demanding settings. For instance, depth cameras, with their ability to function reliably in low-light conditions and provide spatial data independent of ambient lighting, are valuable in the nighttime applications. This indicates that ***only by centering the value of every modality***, we can essentially harness the superiority of all modalities for achieving modality-agnostic segmentation. Therefore, it becomes imperative for the fusion model to distinguish the **robust** and **fragile** modalities, and subsequently incorporate the most robust and fragile modalities to learn a more **resilient** multi-modal framework. The most robust feature is used to **enhance the segmentation accuracy**

while the most fragile feature is incorporated to **reinforce the framework's resilience** against missing modalities.

Contributions: In light of this, we propose an efficient and robust **M**odality-**ag**nost**ic** (**MAGIC**) segmentation framework, that can be flexibly paired with various backbones, ranging from efficient to high-performance models. Our method comprises two plug-and-play modules that enable efficient multi-modal learning and improve the modality-agnostic robustness of segmentation models. Firstly, we introduce a Multi-modal Aggregation Module (MAM) to efficiently process features from multi-modal data and extract complementary scene information from all the modalities without relying on a specific one.

On top, an Arbitrary-modal Selection Module (ASM) is proposed to dynamically fuse the modality-agnostic scene features during training and improve the backbone model's robustness with arbitrary-modal input during inference. Specifically, it utilizes the aggregated features from MAM as a benchmark to rank the multi-modal features based on the similarity scores. It then merges the selected salient features together to obtain predictions, so as to achieve modality-agnostic capabilities. This way, we can eradicate the reliance on RGB modality and better overcome sensor failures while significantly enhancing the segmentation performance, see Fig. 1 (b) and (c). Moreover, our method combines the MAM's prediction and the ground truth to soften the supervision for the predictions of ASM for better convergence and to prevent training instability.

Extensive experiments under the widely considered multi-modal segmentation setting on two multi-modal benchmarks show that our framework outperforms the existing multi-modal semantic segmentation methods (+**1.33%** & +**1.47%**) while reducing model parameters by **60%**. Moreover, we evaluate our method in the novel modality-agnostic settings with arbitrary-modal inputs. The results show that our methods significantly outperforms existing works by a large margin (+**19.41%** & +**12.83%**).

2 Related Work

Semantic Segmentation is a fundamental vision task with many applications, such as autonomous driving [8,9,13,18,19,32,50,51,55,62,75,96–98,101,102,108]. Approaches for semantic segmentation can be categorized according to their basic computing paradigm, namely convolution and self-attention. The fully convolutional networks (FCN) [43] made significant progress in semantic segmentation with their end-to-end pixel-wise classification paradigm. Subsequent works improve the performance by exploring the multi-scale features [10, 11, 23, 93], attention blocks [14,20,26,82], edge cues [3,16,21,34,60], and context priors [24, 39, 81, 85]. More recently, various self-attention-based transformers have been proposed for semantic segmentation [22,41,42,57,63,65,66,76,78,83,91,94,106].

While these works achieve promising performance on RGB images under perfect lighting and motion conditions, they still suffer under extreme scenarios with complex lighting and weather conditions. We incorporate these segmentation models as the backbones, and propose two plug-and-play modules to achieve

robust multi-modality semantic segmentation and modality-agnostic segmentation with arbitrary-modal input data.

Multi-modal Semantic Segmentation has been extensively studied, with a focus on fusing the RGB modality with complementary modalities such as depth [6, 12, 15, 28, 31, 44–46, 56, 61, 69, 70, 80, 103], thermal [7, 27, 38, 52, 54, 73, 77, 89, 90, 105], polarization [30, 48, 74], events [1, 5, 88, 99, 99, 104], and LiDAR [2,33,35,40,68,79,84,109]. With the development of novel sensors, various approaches [36,68–70,86,87,100] have been proposed to scale from dual modality fusion to multiple modality fusion for robust scene understanding abilities, *e.g.*, MCubeSNet [36].

From the architecture design perspective, these methods can be divided into three categories, including merging with separate branches [4, 47, 72, 89], distributing with a joint branch [12,68], and fusion with asymmetric branches [86, 87]. These approaches regard RGB modality as the center and the other modalities as the auxiliary, yielding an asymmetric architecture with two branches. In particular, CMNeXT [87] achieves multi-modal semantic segmentation with arbitrary-modal complements by relying on the RGB branch and treating other modalities as auxiliary inputs. However, the RGB modality may struggle in certain circumstances, *e.g.*, nighttime, thus it is imperative for the fusion model to learn a more resilient multi-modal framework without relying on a specific sensor. Our MAGIC framework treats all visual modalities equally and avoids relying on each modality. This way, our method can eliminate the dependence on RGB modality and better overcome sensor failures while significantly enhancing the segmentation performance.

3 Methodology

In this section, we introduce our **MAGIC** framework. As depicted in Fig. 2, it consists of two pivotal modules: the Multi-modal Aggregation Module (MAM) and the Arbitrary-modal Selection Module (ASM). Our approach takes multiple visual modalities as inputs [3].

3.1 Task Parameterization

Inputs: Our framework takes input data from four modalities, each captured or synthesized within the same scene. Specifically, we consider RGB images $\mathbf{R} \in \mathbb{R}^{h \times w \times 3}$, depth maps $\mathbf{D} \in \mathbb{R}^{h \times w \times C^D}$, LiDAR point clouds $\mathbf{L} \in \mathbb{R}^{h \times w \times C^L}$, and event stack images $\mathbf{E} \in \mathbb{R}^{h \times w \times C^E}$. Here, $C^D = C^L = C^E = 3$. Additionally, we incorporate the corresponding ground truth label $\mathbf{Y}$ spanning K categories. Unlike existing methods that process multi-modal data individually, our method takes a mini-batch $\{r, d, l, e\}$ containing samples from all modalities, where $r \in \mathbf{R}$, $d \in \mathbf{D}$, $l \in \mathbf{L}$, and $e \in \mathbf{E}$.

Outputs: Given the multi-modal data mini-batch $\{r, d, l, e\}$, we feed it into our backbone, producing multi-modal features $\{f_r, f_d, f_l, f_e\}$, as depicted in Fig. 2.

[3] Here, we take the modalities in DELIVER [87] as an example.

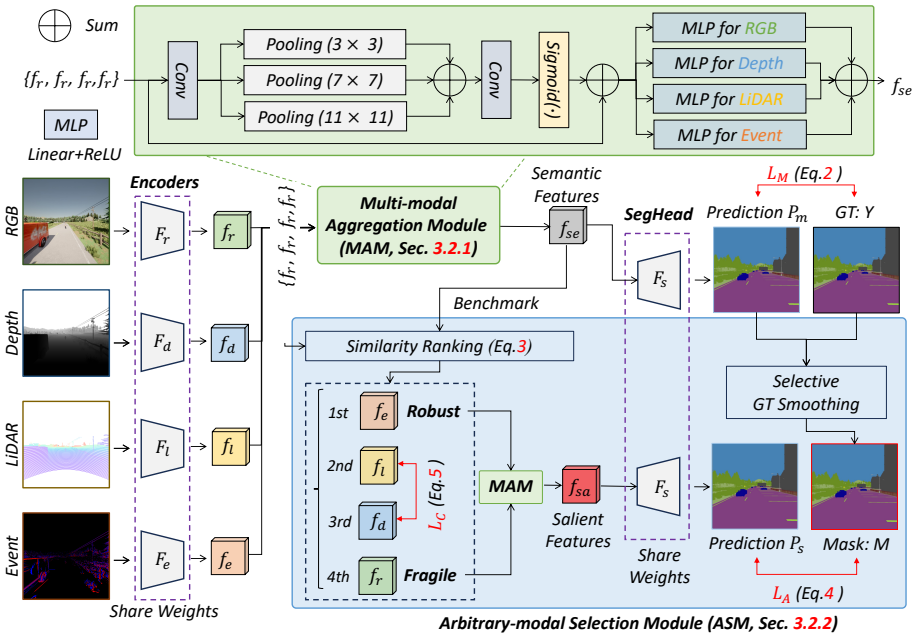

Fig. 2: Overall framework of our MAGIC framework, incorporates plug-and-play multi-modal aggregation and arbitrary-modal selection modules.

Subsequently, these features are concurrently processed by MAM and ASM, yielding semantic feature f_{se} and salient feature f_{sa}, respectively. The segmentation head then leverages f_{se} and f_{sa} to derive the MAM predictions P_m and the ASM predictions P_a, respectively.

3.2 MAGIC Architecture

As depicted in Fig. 2, our MAGIC framework adopts prevailing backbone models, such as SegFormer [76], as the feature encoder and as the segmentation head (SegHead) for each modality. The multi-modal mini-batch $\{r, d, l, e\}$ is directly ingested by weight-shared encoders F_r, F_d, F_l, F_e within the backbone model. This process yields high-level multi-modal features $\{f_r, f_d, f_l, f_e\}$ as:

$$\{f_r, f_d, f_l, f_e\} = F_r(r), F_d(d), F_l(l), F_e(e). \tag{1}$$

3.2.1 Multi-modal Aggregation Module (MAM). Upon acquiring the high-level multi-modal features, MAM is designed to further extract the semantic-rich feature from $\{f_r, f_d, f_l, f_e\}$, paving the way for achieving robust arbitrary-modal capabilities. Our MAM centers the values of every modality and simultaneously extracts complementary features from every modality. As shown in Fig. 2, multi-modal data is formed as a batch rather than concatenated and MLP

layers are assigned to specific modalities. *Note that, when performing modality-agnostic validation with arbitrary input modalities, only the input modalities' corresponding layers are used to give predictions.* This essentially differs from prior methods, *e.g.*, [87], which employs the Self-Query Hub and Parallel Pooling Mixer to prioritize extracted features from auxiliary modalities before fusing them with the RGB feature. As depicted in Fig. 2, the architecture of MAM comprises three main components: a parallel multi-layer perceptron (MLP), a parallel pooling layer, and integrated residual connections.

Specifically, as shown in Fig. 2, $\{f_r, f_d, f_l, f_e\}$ is first processed with a *Conv* layer, then parallel pooling layers (3×3, 7×7, 11×11) are employed to explore spatial information from different scales. These features are then aggregated and processed with another *Conv* layer with Sigmoid activation. Then, parallel MLP layers are utilized to aggregate these multi-modal features with the original $\{f_r, f_d, f_l, f_e\}$. The aggregated feature from outputs of MLP layers, denoted as f_{se}, acts as the semantic representation for generating predictions P_m via the segmentation head, *i.e.*, decoder F_s. The predicted P_m is supervised by the ground truth (GT), $y \in \mathbf{Y}$. Ultimately, the cross-entropy is used as the supervision loss $\mathcal{L}_M$:

$$\mathcal{L}_M = -\sum_{0}^{K-1} Y \cdot log(P_m). \tag{2}$$

3.2.2 Arbitrary-modal Selection Module (ASM). Alongside MAM, we introduce the ASM which is utilized during training, leveraging the most robust feature to *enhance the predictive accuracy* of the framework. The integration of the most fragile features — those which are extracted from the challenging input data samples — serves to *reinforce the framework's resilience against missing modalities* in such challenging scenarios. As illustrated in Fig. 3, our ASM encompasses two principal components: cross-modal semantic similarity ranking and cross-modal semantic consistency training. We now describe the details.

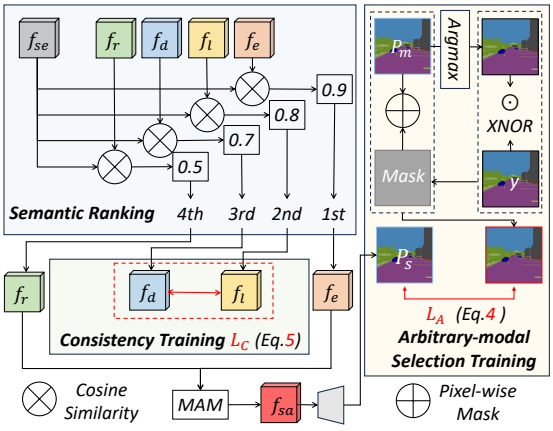

Fig. 3: Illustration of the proposed arbitrary-modal selection module (**ASM**).

Cross-modal Semantic Similarity Ranking. serves as a mechanism to compare the multi-modal features $\{f_r, f_d, f_l, f_e\}$ against the semantic feature

f_{se} derived from MAM. This thereby yields ranked similarity scores. The nature of multi-modal data is inherently diverse, spanning a broad spectrum of conditions. A case in point is the *DELIVER* dataset. As delineated in [87], it features four unique environmental scenarios and documents five episodes of partial sensor malfunctions. Additionally, the intricacies of real-world environments may pose even more heterogeneous challenges. In light of such complexities, it becomes imperative for neural networks to adeptly discern the robust modalities from the fragile ones at the feature level. By integrating both the pinnacle of robustness and the nadir of fragility in modalities, a more resilient multi-modal framework can be cultivated. ASM employs the semantic feature f_{se} from MAM as a benchmark to rank $\{f_r, f_d, f_l, f_e\}$ as:

$$f_{rf}, f_{rm} = Rank(Cos(\{f_r, f_d, f_l, f_e\}, f_{se})), \tag{3}$$

where f_{rf} denotes the features that consist of the top-1 (Robust) and last-1 (Fragile) features in this semantic ranking, signifying the most robust and most fragile modalities, respectively; f_{rm} represents the remaining features in the semantic ranking; *Rank* means sorting from largest to smallest; and $Cos(\cdot)$ is the cosine similarity. The resulting f_{rf} are then passed to another MAM to aggregate the final salient feature for generating predictions P_s with the SegHead.

Beyond the scope of feature-level modality-agnostic learning, our ASM, depicted in Fig. 3, introduces a prediction-level arbitrary-modal selection training strategy, centered on y. It aims to bolster the supervision of predictions P_a by selectively merging predictions P_m with the hard ground truth label, y. Initially, the logits of the predictions P_m undergo an *argmax* operation. These are then amalgamated with the hard label y, culminating in the creation of a mask, denoted as M. This mask M is designed to retain the predicted logits when category predictions from $argmax(P_m)$ coincide with those of y for a given pixel. Conversely, it discards predicted logits in instances of discrepancy between the category predictions. Thereafter, this mask M is harnessed as the supervision signal for P_s. This strategy ensures that the arbitrary-feature predictions receive focused and positive supervision and fosters enhanced convergence and averts potential training instabilities. Overall, the arbitrary-modal selection loss is:

$$\mathcal{L}_S = -\sum_{0}^{K-1} M \cdot log(P_s). \tag{4}$$

Cross-modal Semantic Consistency Training. With the cross-modal semantic similarity ranking, the top-1 and last-1 ranked features are attained. We then ***impose semantic consistency*** training between the remaining features f_{rm}, see Fig. 3. The intuition is that the captured semantics of a scene are identical across modalities because the multi-modal data is captured in the same scenario. Meanwhile, due to the distinct data formats and unique properties of different sensors, it is non-trivial to directly align the remaining features f_{rm} from different modalities. Intuitively, ASM also takes the semantic feature f_{sa} from MAM as the surrogate and implicitly aligns the correlation, *i.e.*, cosine similarity, between the remaining features and the semantic feature. For brevity, we use the abbreviation $c_1 = Cos(f^1_{rm}, f_{sa})$ and $c_2 = Cos(f^2_{rm}, f_{sa})$ to represent

the correlations. The consistency training loss can be formulated as:

$$\mathcal{L}_C = \sum_{0}^{K-1}(c_1 log \frac{c_1}{\frac{1}{2}(c_1+c_2)} + c_2 log \frac{c_2}{\frac{1}{2}(c_1+c_2)}). \tag{5}$$

This implicit alignment makes it better to align the features from the scene-semantic consistency perspective.

Training We train our MAGIC framework by minimizing the total loss $\mathcal{L}$ – a linear combination of the losses of $\mathcal{L}_M$, $\mathcal{L}_S$, and $\mathcal{L}_C$:

$$\mathcal{L} = \mathcal{L}_M + \lambda \mathcal{L}_S + \beta \mathcal{L}_C, \tag{6}$$

where λ and β are hyper-parameters for trade-off. The ASM is only utilized in training while the inference is achieved by the backbone together with our MAM.

4 Experiments

4.1 Datasets and Implementation Details

Datasets 1) DELIVER [87] is a large-scale multi-modal dataset that contains depth, LiDAR, Views, Event, and RGB data, with precise annotations of 25 semantic categories. DELIVER considers both environmental conditions (cloudy, foggy, night, and rainy) and sensor failure cases (motion blur, over-exposure, under-exposure, LiDAR-Jitter, *etc.*) to introduce challenges for robust perception. We follow the official training settings in [87] and convert all data into three channel images. **2) MCubeS [37]** is a material segmentation dataset with 20 categories and pairs of RGB, Near-Infrared (NIR), Degree of Linear Polarization (DoLP), and Angle of Linear Polarization (AoLP) images.

Implementation Details We train our framework on 8 $\times$ NVIDIA GPUs with an initial learning rate of $6e^{-5}$, which is scheduled by the poly strategy with power 0.9 over 200 epochs. The first 10 epochs are to warm-up framework with 0.1 $\times$ the original learning rate. We use AdamW optimizer with epsilon $1e^{-8}$, weight decay $1e^{-2}$, and the batch size is 1 on each GPU. The images are augmented by random resize with ratio 0.5-2.0, random horizontal flipping, random color jitter, random gaussian blur, and random cropping to 1024 $\times$ 1024 on DELIVER [87], while to 512 $\times$ 512 on MCubeS [37]. ImageNet-1K pre-trained weight is used as the pre-trained weight for the backbone model.

Experimental Settings. 1) Multi-modal Semantic Segmentation refers to the process of training and validating models using the same modality inputs, without any absence of training modality data. **2) Modality-agnostic Semantic Segmentation** means we evaluate all the possible combinations of input modalities and average all results to obtain final mean results.

4.2 Experimental Results

Multi-modal Semantic Segmentation: Tab. 1 presents a quantitative comparison between our proposed Magic framework and the SoTA, CMNeXt [87],

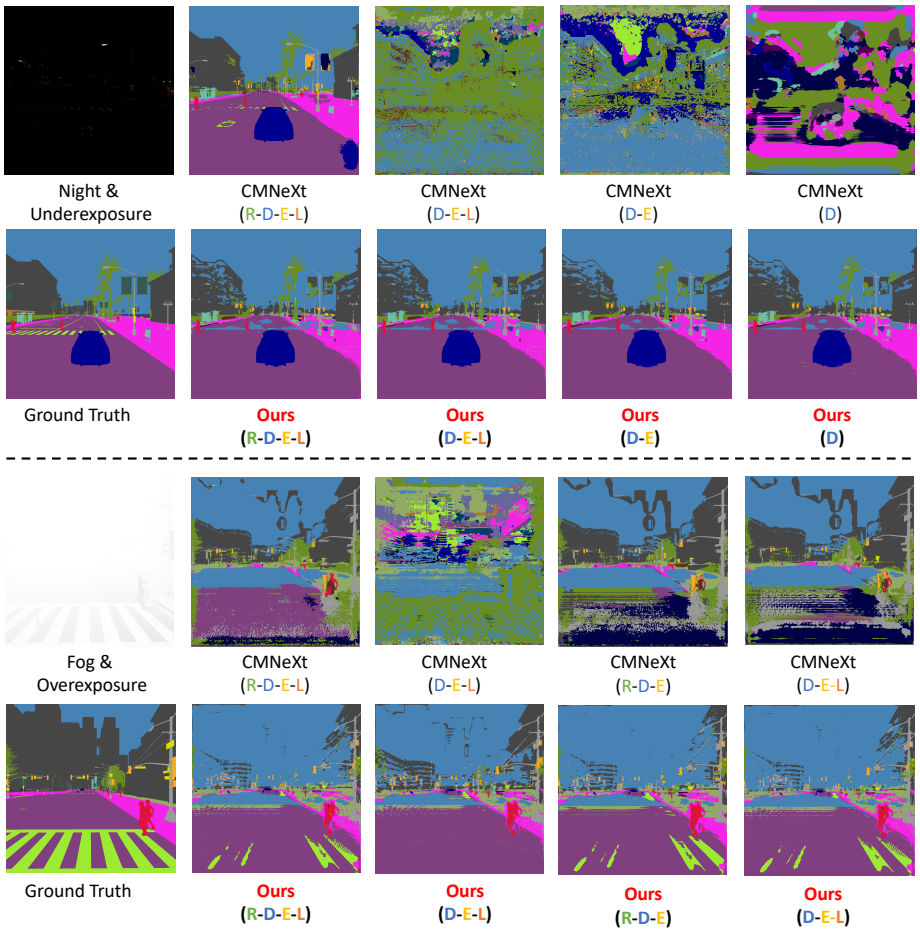

Fig. 4: Visualization of arbitrary inputs using {**R**GB, **D**epth, **E**vent, **L**iDAR} on DELIVER. *(More visualization under different conditions refer to the suppl.)*

on the DELIVER dataset, utilizing four input modalities. Notably, our Magic framework, which contains only 42% of the parameters of CMNeXt (specifically **24.73M** out of 58.73M), surpasses CMNeXt in a majority of categories. For instance, for the 'Fence' category, our model's performance is **49.27%**, improving upon CMNeXt's 43.12% by +**6.15%**. Overall, there's a notable increment of +1.33% in mIoU. These results underscore the effectiveness of our multi-modal input batch approach and the proposed plug-and-play modules, positioning them as advancements over the traditional Hub2Fuse, separate branch, and joint branch paradigms (*more comparison refer to the suppl.*).

In Tab. 1, we present results not only for the DELIVER dataset, which includes RGB, Depth, Event, and LiDAR modalities, but also for the MCubeS [37]

Table 1: Per-class results on DELIVER [87] and MCubeS [37] datasets. Abbreviations: M.: Method; B.b.: Backbone; #P(M): #Param(M); Seg-B2: SegFormer-B2. (*Selected categories in this table, all categories refer to the suppl.*)

DELIVER [87]

Method	Backbone	#Param (M)	Building	Fence	Pedestrain	Road	Sidewalk	Cars	Wall	
CMNeXt	Seg-B2	58.73	89.41	43.12	76.51	98.18	82.27	84.98	69.39	Mean
Ours	Seg-B2	24.73	89.66	49.27	76.54	98.40	86.22	90.94	70.86	

Method	Backbone	#Param (M)	Traffic Sign	Ground	Bridge	Groundrail	Static	Water	Terrain	
CMNeXt	Seg-B2	58.73	70.57	1.31	53.61	55.01	33.58	23.96	83.94	66.30
Ours	Seg-B2	24.73	72.88	2.62	59.28	73.08	35.70	30.93	84.00	**67.66**

MCubeS [37]

Method	Backbone	#Param (M)	Asphalt	Concrete	Roadmarking	Fabirc	Glass	Plaster	Rubber
CMNeXt	Seg-B2	58.73	84.71	45.23	74.80	32.18	54.34	0.69	28.54
Ours	Seg-B2	24.73	88.86	50.79	74.98	38.55	55.27	0.86	34.21

Method	Backbone	#Param (M)	Traffic Sign	Ceramic	Cobblestone	Brick	Water	Sky	Mean
CMNeXt	Seg-B2	58.73	28.72	26.81	68.67	43.19	54.35	96.52	51.54
Ours	Seg-B2	24.73	31.98	31.21	71.67	46.47	54.72	96.60	**53.01**

dataset. The latter comprises four modalities: Image, AoLP, DoLP, and NIR. Notably, our Magic framework consistently surpasses the state-of-the-art CMNeXt [87] by an impressive +**1.47** mIoU. It also achieves superior results in a majority of categories. For instance, for the 'Concrete' category, our model registers **50.79%**, outdoing CMNeXt's 45.23% by +**5.56%**. These results emphasize the efficacy of our Magic framework, and they highlight that our proposed MAM and ASM function as efficient plug-and-play modules tailored for multi-modal learning (*For results of training with two modalities, please refer the to suppl.*).

Modality-agnostic Semantic Segmentation Unlike the approach in [87] where arbitrary modality inputs necessitate the inclusion of RGB data, our methodology operates effectively on arbitrary inputs without specifically relying on any given modality. To evaluate the resilience of our approach with such arbitrary modality inputs, we apply our Magic framework to both the DELIVER and MCubeS [37] datasets. As shown in Tab. 2, models are trained with the total four modalities and is validated with arbitrary modality combination. Our Magic with SegFormer-B2 significantly outperforms the CMNeXt [87] at nearly all the validation scenarios and achieves +19.41% and +14.97% mIoU performance gain than CMNeXt [87] at DELIVER [87] and MCubeS [37] datasets, respectively. Especially for the RGB data absence scenarios, our Magic with SegFormer-B2 eclipses much more performance, such as Depth only (**57.59%** vs. 0.81% → +**56.78%**), Depth with Event (**57.62%** vs. 21.48% → +**36.14%**), and Depth with LiDAR (**57.60%** vs. 3.83 → +**53.77%**). Importantly, our Magic with SegFormer-B2 only has 42% parameters of CMNeXt. Furthermore, our Magic with SegFormer-B0 even surpasses CMNeXt [87] by +**15.24** and +**14.97** mIoU with only **6%** parameters. All these results show that our MAM and ASM are powerful plug-and-play modules for multi-modal visual learning, especially for the modality-agnostic segmentation scenarios with arbitrary-modal inputs.

Table 2: Validation with arbitrary-modal inputs. All methods are trained with four modalities, and the metric is mIoU for all numbers. Abbreviations: Seg-B0: SegFormer-B0, R: RGB; D: Depth; E: Event; L: LiDAR; I: Image; A: Aolp; D: Dolp; N: Nir.

	Method	Backbone	#Param (M)	R	D	E	L	RD	RE	RL	DE	
DELIVER [87]	CMNeXt	Seg-B2	58.73	3.76	0.81	**1.00**	**0.72**	50.33	13.23	18.22	21.48	
	MAGIC	Seg-B0	**3.72**	32.60	55.06	0.52	0.39	63.32	33.02	33.12	55.16	
		Seg-B2	24.73	**41.97**	**57.59**	0.40	0.37	**67.65**	**41.93**	**42.00**	**57.62**	
	Method	Backbone	#Param (M)	DL	EL	RDE	RDL	REL	DEL	RDEL	Mean	
	CMNeXt	Seg-B2	58.73	3.83	**2.86**	66.24	66.43	15.75	46.29	66.30	25.25	
	MAGIC	Seg-B0	**3.72**	55.17	0.26	63.37	63.36	33.32	55.26	63.40	40.49	+15.24
		Seg-B2	24.73	**57.60**	0.27	**67.66**	**67.65**	**41.93**	**57.63**	**67.66**	**44.66**	+19.41
	Method	Backbone	#Param (M)	I	A	D	N	IA	ID	IN	AD	
MCubeS [37]	CMNeXt	Seg-B2	58.73	1.86	**1.54**	2.51	2.28	47.96	43.67	45.90	**6.99**	
	MAGIC	Seg-B0	**3.72**	46.35	0.67	33.33	1.00	47.53	47.26	45.88	0.55	
		Seg-B2	24.73	**51.91**	0.32	**34.52**	**2.66**	**52.24**	**52.16**	**52.57**	1.98	
	Method	Backbone	#Param (M)	AN	DN	IAD	IAN	ID+N	ADN	IADN	Mean	
	CMNeXt	Seg-B2	58.73	7.58	9.95	50.07	48.77	48.83	8.06	51.54	25.17	
	MAGIC	Seg-B0	**3.72**	35.32	35.70	47.85	47.01	46.71	35.64	47.58	34.56	+9.39
		Seg-B2	24.73	**36.01**	**37.09**	**52.48**	**52.82**	**52.73**	**37.57**	**53.01**	**38.00**	+14.97

In Tab. 3, we show the modality-agnostic validation results of training models with 3 modalities on both datasets. Our Magic consistently outperforms CMNeXt [87] at all arbitrary modality inputs and achieve more performance gains. For instance, our Magic with SegFormer-B2 achieves +**32.44%** and +**28.35%** mIoU performance gains on DELIVER [87], +**24.53%** and +**22.22%** mIoU performance boost on MCubeS [37] with only **42%** parameters. This further confirms the superiority and robustness of our proposed Magic framework in the arbitrary input setting. In Fig. 4, we present a comparison of the segmentation results obtained using Magic and CMNeXt [87]. The results demonstrate that our Magic consistently performs well with arbitrary inputs, whereas CMNeXt is fragile in most scenarios. Notably, our Magic does not rely on a specific modality and is relatively insensitive to the absence of modalities, which further enhances the robustness of full scene segmentation under varying lighting and weather conditions, such as cloudy, rain, and motion blur.

Effectiveness of the Loss Functions. To evaluate the effectiveness of the proposed loss functions, we conduct ablation experiments on two datasets, DELIVER and MCubeS, using SegFormer-B0 and -B2 backbones. As delineated in Tab. 4, each of our proposed modules and associated loss functions consistently enhances the performance of multi-modal semantic segmentation. Significantly, our MAM in conjunction with $\mathcal{L}_M$ results in mIoU improvements of +**8.21%** and +**5.51%** for SegFormer-B0 and -B2, respectively. Building upon our MAM,

Table 3: Results of modality-agnostic validation with three modalities.

Method	Backbone	Training	DELIVER dataset							Mean	$\Delta\uparrow$
			R	D	L	RD	RL	DL	RDL		
CMNeXt [87]	Seg-B2	RDL	1.87	1.87	2.01	52.90	23.35	4.67	65.50	21.74	-
MAGIC(ours)	Seg-B0	RDL	32.41	56.20	1.40	62.64	32.61	**56.29**	62.64	43.46	+21.72
	Seg-B2	RDL	**37.08**	**60.52**	**2.38**	**67.66**	**67.62**	37.36	**67.63**	**54.18**	**+32.44**
			R	D	E	RD	RE	DE	RDE	Mean	$\Delta\uparrow$
CMNeXt [87]	Seg-B2	RDE	1.75	1.71	2.06	53.68	9.66	2.84	64.44	19.45	-
MAGIC(ours)	Seg-B0	RDE	32.96	55.90	2.15	62.52	33.25	56.00	62.49	43.61	+24.16
	Seg-B2	RDE	**38.13**	**60.42**	**2.75**	**67.16**	**38.12**	**60.87**	**67.16**	**47.80**	**+28.35**
Method	Backbone	Training	McubeS dataset							Mean	$\Delta\uparrow$
			I	A	N	IA	IN	AN	IAN		
CMNeXt [87]	Seg-B2	IAN	2.16	2.09	2.45	41.46	44.66	10.73	47.96	21.64	-
MAGIC(ours)	Seg-B0	IAN	47.04	0.33	33.00	47.70	46.17	34.92	47.43	36.66	+15.02
	Seg-B2	IAN	**50.36**	**31.25**	**43.69**	**50.79**	**50.35**	**45.92**	**50.80**	**46.17**	**+24.53**
			I	A	D	IA	ID	AD	IAD	Mean	$\Delta\uparrow$
CMNeXt [87]	Seg-B2	IAD	1.15	1.87	1.27	47.14	47.80	12.52	49.48	23.03	-
MAGIC(ours)	Seg-B0	IAD	46.58	0.01	4.84	46.68	47.51	3.87	47.78	28.18	+5.15
	Seg-B2	IAD	**49.05**	**35.45**	**39.14**	**50.52**	**50.20**	**41.60**	**50.79**	**45.25**	**+22.22**

the ASM and its associated $\mathcal{L}_A$ further register mIoU increments of $+$**10.78%** and $+$**8.12%** across the dual backbones. Complementing these modules, the consistency training loss $\mathcal{L}_C$ yields $+$**11.99%** and $+$**8.99%** mIoU advancements relative to the baseline.

Ablation Study of MAM Components. As indicated in Fig. 5 (c), we evaluate individual components within our proposed MAM. Removing any of the components invariably results in diminished performance. This underscores the indispensable contribution of each component to the efficacy of MAM.

Ablation of Pooling Size. We conduct an ablation study focusing on the pooling size used in the parallel pooling of MAM on MCubeS, as depicted in Fig. 5 (d). Our empirical findings indicate that a pooling size of (3,7,11) consistently delivers the optimal mIoU performance.

Selection of the Salient Features in ASM. The process of selecting salient features is pivotal for ASM. In order to evaluate the efficacy of this selection strategy, we organize the features based on their congruence with the semantic features, subsequently marking the salient features with stars, as shown in Fig. 6 (a). Our results underscore that choosing features from both extremes of the sorted sequence optimally supports modality-agnostic segmentation, resonating with our initial hypothesis. We also conduct experiments by randomly

Table 4: Ablation study of different loss function combinations.

Backbone	Loss			DELIVER [87]				MCubeS [37]|			
	$\mathcal{L}_M$	$\mathcal{L}_S$	$\mathcal{L}_C$	mIoU	$\Delta\uparrow$	Acc	$\Delta\uparrow$	mIoU	$\Delta\uparrow$	Acc	$\Delta\uparrow$
Seg-B0	-	-	-	51.41	-	59.76	-	42.43	-	51.34	-
	✓	-	-	59.62	+8.21	68.42	+8.66	42.91	+0.48	52.72	+1.38
	✓	✓	-	62.19	+10.78	70.12	+10.36	46.50	+4.07	56.11	+4.77
	✓	✓	✓	63.40	+11.99	70.84	+11.08	47.58	+5.15	56.35	+5.01
Seg-B2	-	-	-	58.67	-	65.48	-	46.51	-	56.02	-
	✓	-	-	64.18	+5.51	73.35	+7.87	49.66	+1.80	58.67	+1.03
	✓	✓	-	66.79	+8.12	74.08	+8.60	52.24	+4.38	61.70	+4.06
	✓	✓	✓	67.66	+8.99	74.69	+9.21	53.01	+5.15	62.87	+5.23

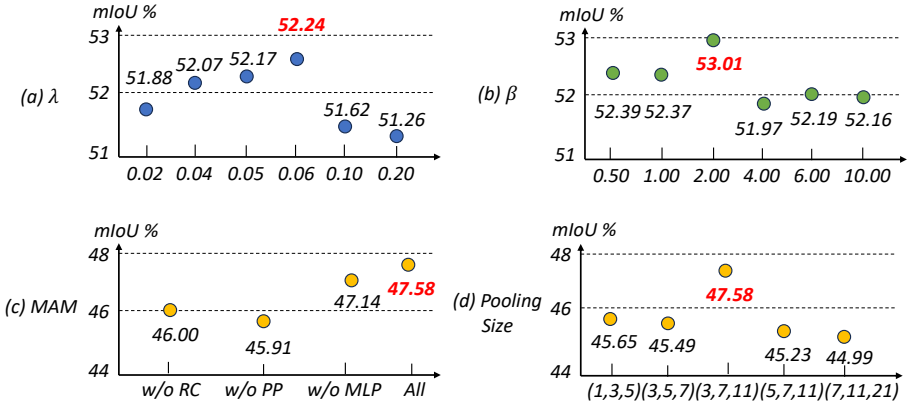

Fig. 5: (a),(b): Ablation of λ and β on MCubeS [37]. (c),(d): Ablation of components in MAM with MiT-B0 on MCubeS [37].

dropping one or a few modalities with only MAM at train time, the result is 42.44 mIoU which is less than ours 44.66 mIoU.

5 Ablation Study and Analysis

Ablations of λ and β. Fig. 5 (a) and (b) present the ablation results for varying values of hyper-parameters λ and β.

6 Discussion

Every Modality Matters. Contrary to the assumptions in prior works such as [87], which advocate for the indispensability of RGB representation in semantic segmentation, we propose that every modality holds significance. Our

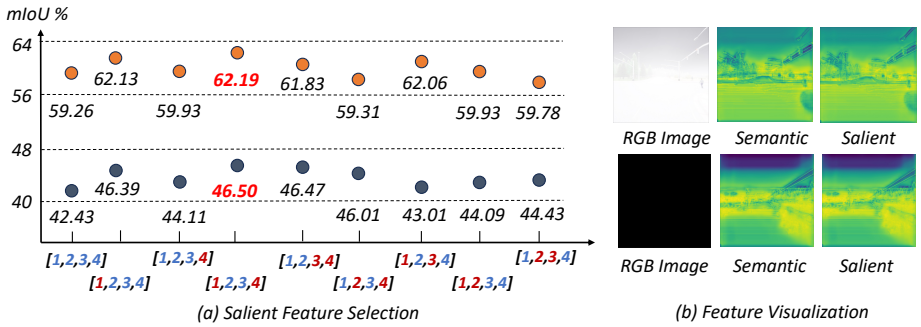

Fig. 6: (a) Ablation study on the salient feature selection strategy, where [1,2,3,4] stand for the ranked multi-modal features and the **blue** numbers denotes unselected features, and **red** numbers represents the selected salient features among the ranked multi-modal features. (b) Visualization of both semantic and salient features.

contention is that every modality brings value and should not be overlooked. As evidenced in Tab. 2, introducing each modality during inference within our Magic model corresponds to enhancements in mIoU; for instance, transitioning from R+D+L to R+D+E+L improves from 67.65% to 67.66%.

Visualization of Salient Features. In Fig. 6 (b), we provide visualization of the RGB features, semantic features derived from MAM, and the salient features extracted by ASM. Notably, both the semantic and salient features exhibit a richer capture of scene details than the RGB features. This underscores the efficacy of our MAM and ASM in harnessing the potential of multi-modal input.

7 Conclusion

In this paper, we presented our proposed Magic framework for modality-agnostic semantic segmentation, which can be implemented with various existing segmentation backbones. We introduced a multi-modal aggregation module (MAM) for extracting complementary scene information and a unified arbitrary-modal selection module (ASM) for enhancing the backbone model's robustness for arbitrary-modal input data. Our Magic significantly outperformed previous multi-modal methods in both multi-modal and modality-agnostic semantic segmentation benchmarks by a large margin, achieving new state-of-the-art performance.

Limitations and Future Directions. Despite its notable contributions, MAGIC exhibits sub-optimal performance due to inherent data characteristics in a few scenarios. Future efforts will focus on refining and enhancing our modality-agnostic plug-and-play modules to ensure consistent and improved performance.

8 Acknowledgement

This paper is supported by the National Natural Science Foundation of China (NSF) under Grant No. NSFC22FYT45, the Guangzhou City, University and Enterprise Joint Fund under Grant No.SL2022A03J01278, and Guangzhou Fundamental and Applied Basic Research (Grant Number: 2024A04J4072)

References

1. Alonso, I., Murillo, A.C.: Ev-segnet: Semantic segmentation for event-based cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
2. Borse, S., Klingner, M., Kumar, V.R., Cai, H., Almuzairee, A., Yogamani, S., Porikli, F.: X-align: Cross-modal cross-view alignment for bird's-eye-view segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3287–3297 (2023)
3. Borse, S., Wang, Y., Zhang, Y., Porikli, F.: Inverseform: A loss function for structured boundary-aware segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5901–5911 (2021)
4. Broedermann, T., Sakaridis, C., Dai, D., Van Gool, L.: Hrfuser: A multi-resolution sensor fusion architecture for 2d object detection. arXiv preprint arXiv:2206.15157 (2022)
5. Cao, J., Zheng, X., Lyu, Y., Wang, J., Xu, R., Wang, L.: Chasing day and night: Towards robust and efficient all-day object detection guided by an event camera. arXiv preprint arXiv:2309.09297 (2023)
6. Cao, J., Leng, H., Lischinski, D., Cohen-Or, D., Tu, C., Li, Y.: Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7088–7097 (2021)
7. Chen, G., Shao, F., Chai, X., Chen, H., Jiang, Q., Meng, X., Ho, Y.S.: Modality-induced transfer-fusion network for rgb-d and rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **33**(4), 1787–1801 (2022)
8. Chen, J., Deguchi, D., Zhang, C., Zheng, X., Murase, H.: Clip is also a good teacher: A new learning framework for inductive zero-shot semantic segmentation. arXiv preprint arXiv:2310.02296 (2023)
9. Chen, J., Deguchi, D., Zhang, C., Zheng, X., Murase, H.: Frozen is better than learning: A new design of prototype-based classifier for semantic segmentation. Pattern Recognition **152**, 110431 (2024)
10. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
11. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
12. Chen, L.Z., Lin, Z., Wang, Z., Yang, Y.L., Cheng, M.M.: Spatial information guided convolution for real-time rgbd semantic segmentation. IEEE Transactions on Image Processing **30**, 2313–2324 (2021)

13. Cheng, H.X., Han, X.F., Xiao, G.Q.: Cenet: Toward concise and efficient lidar semantic segmentation for autonomous driving. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 01–06. IEEE (2022)
14. Choi, S., Kim, J.T., Choo, J.: Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9373–9383 (2020)
15. Cong, R., Lin, Q., Zhang, C., Li, C., Cao, X., Huang, Q., Zhao, Y.: Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection. IEEE Transactions on Image Processing **31**, 6800–6815 (2022)
16. Ding, H., Jiang, X., Liu, A.Q., Thalmann, N.M., Wang, G.: Boundary-aware feature propagation for scene segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6819–6829 (2019)
17. Duan, S., Shi, Q., Wu, J.: Multimodal sensors and ml-based data fusion for advanced robots. Advanced Intelligent Systems **4**(12), 2200213 (2022)
18. Fantauzzo, L., Fanì, E., Caldarola, D., Tavera, A., Cermelli, F., Ciccone, M., Caputo, B.: Feddrive: Generalizing federated learning to semantic segmentation in autonomous driving. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 11504–11511. IEEE (2022)
19. Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., Dietmayer, K.: Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems **22**(3), 1341–1360 (2020)
20. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3146–3154 (2019)
21. Gong, J., Xu, J., Tan, X., Zhou, J., Qu, Y., Xie, Y., Ma, L.: Boundary-aware geometric encoding for semantic segmentation of point clouds. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1424–1432 (2021)
22. Gu, J., Kwon, H., Wang, D., Ye, W., Li, M., Chen, Y.H., Lai, L., Chandra, V., Pan, D.Z.: Multi-scale high-resolution vision transformer for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12094–12103 (2022)
23. Hou, Q., Zhang, L., Cheng, M.M., Feng, J.: Strip pooling: Rethinking spatial pooling for scene parsing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4003–4012 (2020)
24. Hu, X., Yang, K., Fei, L., Wang, K.: Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1440–1444. IEEE (2019)
25. Huang, K., Shi, B., Li, X., Li, X., Huang, S., Li, Y.: Multi-modal sensor fusion for auto driving perception: A survey. arXiv preprint arXiv:2202.02703 (2022)
26. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 603–612 (2019)
27. Hui, T., Xun, Z., Peng, F., Huang, J., Wei, X., Wei, X., Dai, J., Han, J., Liu, S.: Bridging search region interaction with template for rgb-t tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13630–13639 (2023)
28. Ji, W., Yan, G., Li, J., Piao, Y., Yao, S., Zhang, M., Cheng, L., Lu, H.: Dmra: Depth-induced multi-scale recurrent attention network for rgb-d saliency detection. IEEE Transactions on Image Processing **31**, 2321–2336 (2022)

29. Jia, Z., You, K., He, W., Tian, Y., Feng, Y., Wang, Y., Jia, X., Lou, Y., Zhang, J., Li, G., et al.: Event-based semantic segmentation with posterior attention. IEEE Transactions on Image Processing **32**, 1829–1842 (2023)
30. Kalra, A., Taamazyan, V., Rao, S.K., Venkataraman, K., Raskar, R., Kadambi, A.: Deep polarization cues for transparent object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8602–8611 (2020)
31. Lee, M., Park, C., Cho, S., Lee, S.: Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In: European Conference on Computer Vision. pp. 630–647. Springer (2022)
32. Li, J., Dai, H., Ding, Y.: Self-distillation for robust lidar semantic segmentation in autonomous driving. In: European Conference on Computer Vision. pp. 659–676. Springer (2022)
33. Li, J., Dai, H., Han, H., Ding, Y.: Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21694–21704 (2023)
34. Li, X., Li, X., Zhang, L., Cheng, G., Shi, J., Lin, Z., Tan, S., Tong, Y.: Improving semantic segmentation via decoupled body and edge supervision. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. pp. 435–452. Springer (2020)
35. Li, Y., Yu, A.W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Lu, Y., Zhou, D., Le, Q.V., et al.: Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17182–17191 (2022)
36. Liang, Y., Wakaki, R., Nobuhara, S., Nishino, K.: Multimodal material segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19800–19808 (2022)
37. Liang, Y., Wakaki, R., Nobuhara, S., Nishino, K.: Multimodal material segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19800–19808 (June 2022)
38. Liao, G., Gao, W., Li, G., Wang, J., Kwong, S.: Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(11), 7646–7661 (2022)
39. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1925–1934 (2017)
40. Liu, H., Lu, T., Xu, Y., Liu, J., Li, W., Chen, L.: Camliflow: Bidirectional camera-lidar fusion for joint optical flow and scene flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5791–5801 (2022)
41. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022)
42. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
43. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)

44. Lyu, Y., Zheng, X., Kim, D., Wang, L.: Omnibind: Teach to build unequal-scale modality interaction for omni-bind of all. arXiv preprint arXiv:2405.16108 (2024)
45. Lyu, Y., Zheng, X., Wang, L.: Image anything: Towards reasoning-coherent and training-free multi-modal image generation. arXiv preprint arXiv:2401.17664 (2024)
46. Lyu, Y., Zheng, X., Zhou, J., Wang, L.: Unibind: Llm-augmented unified and balanced representation space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26752–26762 (2024)
47. Man, Y., Gui, L.Y., Wang, Y.X.: Bev-guided multi-modality fusion for driving perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21960–21969 (2023)
48. Mei, H., Dong, B., Dong, W., Yang, J., Baek, S.H., Heide, F., Peers, P., Wei, X., Yang, X.: Glass segmentation using intensity and spectral polarization cues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12622–12631 (2022)
49. Milioto, A., Vizzo, I., Behley, J., Stachniss, C.: Rangenet++: Fast and accurate lidar semantic segmentation. In: 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 4213–4220. IEEE (2019)
50. Muhammad, K., Hussain, T., Ullah, H., Del Ser, J., Rezaei, M., Kumar, N., Hijji, M., Bellavista, P., de Albuquerque, V.H.C.: Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. IEEE Transactions on Intelligent Transportation Systems (2022)
51. Nesti, F., Rossolini, G., Nair, S., Biondi, A., Buttazzo, G.: Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2280–2289 (2022)
52. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Caver: Cross-modal view-mixed transformer for bi-modal salient object detection. IEEE Transactions on Image Processing **32**, 892–904 (2023)
53. Park, S.J., Hong, K.S., Lee, S.: Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 4980–4989 (2017)
54. Shivakumar, S.S., Rodrigues, N., Zhou, A., Miller, I.D., Kumar, V., Taylor, C.J.: Pst900: Rgb-thermal calibration, dataset and segmentation network. In: 2020 IEEE international conference on robotics and automation (ICRA). pp. 9441–9447. IEEE (2020)
55. Siam, M., Gamal, M., Abdel-Razek, M., Yogamani, S., Jagersand, M., Zhang, H.: A comparative study of real-time semantic segmentation for autonomous driving. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 587–597 (2018)
56. Song, M., Song, W., Yang, G., Chen, C.: Improving rgb-d salient object detection via modality-aware decoder. IEEE Transactions on Image Processing **31**, 6124–6138 (2022)
57. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7262–7272 (2021)
58. Su, H., Qi, W., Chen, J., Yang, C., Sandoval, J., Laribi, M.A.: Recent advancements in multimodal human–robot interaction. Frontiers in Neurorobotics **17**, 1084000 (2023)

59. Sun, Y., Zuo, W., Liu, M.: Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. IEEE Robotics and Automation Letters **4**(3), 2576–2583 (2019)
60. Takikawa, T., Acuna, D., Jampani, V., Fidler, S.: Gated-scnn: Gated shape cnns for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5229–5238 (2019)
61. Wang, F., Pan, J., Xu, S., Tang, J.: Learning discriminative cross-modality features for rgb-d saliency detection. IEEE Transactions on Image Processing **31**, 1285–1297 (2022)
62. Wang, H., Chen, Y., Cai, Y., Chen, L., Li, Y., Sotelo, M.A., Li, Z.: Sfnet-n: An improved sfnet algorithm for semantic segmentation of low-light autonomous driving road scenes. IEEE Transactions on Intelligent Transportation Systems **23**(11), 21405–21417 (2022)
63. Wang, J., Gou, C., Wu, Q., Feng, H., Han, J., Ding, E., Wang, J.: Rtformer: Efficient design for real-time semantic segmentation with transformer. Advances in Neural Information Processing Systems **35**, 7423–7436 (2022)
64. Wang, J., Wang, Z., Tao, D., See, S., Wang, G.: Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 664–679. Springer (2016)
65. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021)
66. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)
67. Wang, Y.: Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **17**(1s), 1–25 (2021)
68. Wang, Y., Chen, X., Cao, L., Huang, W., Sun, F., Wang, Y.: Multimodal token fusion for vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12186–12195 (2022)
69. Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., Huang, J.: Deep multimodal fusion by channel exchanging. Advances in neural information processing systems **33**, 4835–4845 (2020)
70. Wang, Y., Sun, F., Lu, M., Yao, A.: Learning deep multimodal feature representation with asymmetric multi-layer fusion. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 3902–3910 (2020)
71. Wang, Y., Mao, Q., Zhu, H., Deng, J., Zhang, Y., Ji, J., Li, H., Zhang, Y.: Multimodal 3d object detection in autonomous driving: a survey. International Journal of Computer Vision pp. 1–31 (2023)
72. Wei, S., Luo, C., Luo, Y.: Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20039–20049 (2023)
73. Wu, W., Chu, T., Liu, Q.: Complementarity-aware cross-modal feature fusion network for rgb-t semantic segmentation. Pattern Recognition **131**, 108881 (2022)
74. Xiang, K., Yang, K., Wang, K.: Polarization-driven semantic segmentation via efficient attention-bridged fusion. Optics Express **29**(4), 4802–4820 (2021)

75. Xiao, X., Zhao, Y., Zhang, F., Luo, B., Yu, L., Chen, B., Yang, C.: Baseg: Boundary aware semantic segmentation for autonomous driving. Neural Networks **157**, 460–470 (2023)
76. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
77. Xie, Z., Shao, F., Chen, G., Chen, H., Jiang, Q., Meng, X., Ho, Y.S.: Cross-modality double bidirectional interaction and fusion network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology (2023)
78. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D.: Multi-class token transformer for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4310–4319 (2022)
79. Yan, X., Gao, J., Zheng, C., Zheng, C., Zhang, R., Cui, S., Li, Z.: 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In: European Conference on Computer Vision. pp. 677–695. Springer (2022)
80. Ying, X., Chuah, M.C.: Uctnet: Uncertainty-aware cross-modal transformer network for indoor rgb-d semantic segmentation. In: European Conference on Computer Vision. pp. 20–37. Springer (2022)
81. Yu, C., Wang, J., Gao, C., Yu, G., Shen, C., Sang, N.: Context prior for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12416–12425 (2020)
82. Yuan, Y., Huang, L., Guo, J., Zhang, C., Chen, X., Wang, J.: Ocnet: Object context for semantic segmentation. International Journal of Computer Vision **129**(8), 2375–2398 (2021)
83. Zhang, B., Tian, Z., Tang, Q., Chu, X., Wei, X., Shen, C., et al.: Segvit: Semantic segmentation with plain vision transformers. Advances in Neural Information Processing Systems **35**, 4971–4982 (2022)
84. Zhang, B., Wang, Z., Ling, Y., Guan, Y., Zhang, S., Li, W.: Mx2m: Masked cross-modality modeling in domain adaptation for 3d semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3401–3409 (2023)
85. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7151–7160 (2018)
86. Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelhagen, R.: Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. arXiv preprint arXiv:2203.04838 (2022)
87. Zhang, J., Liu, R., Shi, H., Yang, K., Reiß, S., Peng, K., Fu, H., Wang, K., Stiefelhagen, R.: Delivering arbitrary-modal semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1136–1147 (2023)
88. Zhang, J., Yang, K., Stiefelhagen, R.: Issafe: Improving semantic segmentation in accidents by fusing event-based data. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1132–1139. IEEE (2021)
89. Zhang, Q., Zhao, S., Luo, Y., Zhang, D., Huang, N., Han, J.: Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2633–2642 (2021)

90. Zhang, T., Guo, H., Jiao, Q., Zhang, Q., Han, J.: Efficient rgb-t tracking via cross-modality distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5404–5413 (2023)
91. Zhang, W., Huang, Z., Luo, G., Chen, T., Wang, X., Liu, W., Yu, G., Shen, C.: Topformer: Token pyramid transformer for mobile semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12083–12093 (2022)
92. Zhang, Y., Zhou, Z., David, P., Yue, X., Xi, Z., Gong, B., Foroosh, H.: Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9601–9610 (2020)
93. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
94. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021)
95. Zheng, X., Liu, Y., Lu, Y., Hua, T., Pan, T., Zhang, W., Tao, D., Wang, L.: Deep learning for event-based vision: A comprehensive survey and benchmarks. arXiv preprint arXiv:2302.08890 (2023)
96. Zheng, X., Luo, Y., Wang, H., Fu, C., Wang, L.: Transformer-cnn cohort: Semi-supervised semantic segmentation by the best of both students. arXiv preprint arXiv:2209.02178 (2022)
97. Zheng, X., Luo, Y., Zhou, P., Wang, L.: Distilling efficient vision transformers from cnns for semantic segmentation. arXiv preprint arXiv:2310.07265 (2023)
98. Zheng, X., Pan, T., Luo, Y., Wang, L.: Look at the neighbor: Distortion-aware unsupervised domain adaptation for panoramic semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18687–18698 (2023)
99. Zheng, X., Wang, L.: Eventdance: Unsupervised source-free cross-modal adaptation for event-based object recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17448–17458 (2024)
100. Zheng, X., Zhou, P., Vasilakos, A.V., Wang, L.: 360sfuda++: Towards source-free uda for panoramic segmentation by learning reliable category prototypes. arXiv preprint arXiv:2404.16501 (2024)
101. Zheng, X., Zhou, P., Vasilakos, A.V., Wang, L.: Semantics distortion and style matter: Towards source-free uda for panoramic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27885–27895 (2024)
102. Zheng, X., Zhu, J., Liu, Y., Cao, Z., Fu, C., Wang, L.: Both style and distortion matter: Dual-path unsupervised domain adaptation for panoramic semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1285–1295 (2023)
103. Zhou, H., Qi, L., Wan, Z., Huang, H., Yang, X.: Rgb-d co-attention network for semantic segmentation. In: Proceedings of the Asian conference on computer vision (2020)
104. Zhou, J., Zheng, X., Lyu, Y., Wang, L.: Exact: Language-guided conceptual reasoning and uncertainty estimation for event-based action recognition and more. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18633–18643 (2024)

105. Zhou, W., Zhang, H., Yan, W., Lin, W.: Mmsmcnet: Modal memory sharing and morphological complementary networks for rgb-t urban scene semantic segmentation. IEEE Transactions on Circuits and Systems for Video Technology (2023)
106. Zhu, F., Zhu, Y., Zhang, L., Wu, C., Fu, Y., Li, M.: A unified efficient pyramid transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2667–2677 (2021)
107. Zhu, J., Lai, S., Chen, X., Wang, D., Lu, H.: Visual prompt multi-modal tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9516–9526 (2023)
108. Zhu, J., Luo, Y., Zheng, X., Wang, H., Wang, L.: A good student is cooperative and reliable: Cnn-transformer collaborative learning for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11720–11730 (2023)
109. Zhuang, Z., Li, R., Jia, K., Wang, Q., Li, Y., Tan, M.: Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16280–16290 (2021)