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

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. Due to the spatial constraints within the main paper, this supplementary material provides an expansive elucidation of the methodology proposed and additional experimental evidence supporting its efficacy. Sec. 1 offers a comprehensive examination of the implementation details, providing details into the practical aspects of our MAGIC framework. Sec. 2 presents a broader spectrum of experimental results, emphasizing both quantitative metrics and qualitative assessments. Sec. 3 shows comprehensive ablation studies to evaluate the robustness and contribution of individual components within the MAGIC framework. Finall, Sec. 4 expounds upon the algorithmic foundations of the MAGIC framework, detailing its conceptual and computational structure.

1 Implementation Details

1.1 Datasets

DELIVER [11] is a large-scale multi-modal segmentation dataset which includes Depth, LiDAR, Views, Event, RGB data, based on the CARLA simulator. DELIVER [11] provides cases in two-fold, including four environmental conditions and five partial sensor failure cases. For environmental conditions, there are cloudy, foggy, night, and rainy weather conditions as well as the sunny days. The environmental conditions cause variations in the position and illumination of the sun, atmospheric diffuse reflections, precipitation, and shading of the scene, introducing challenges for robust perception. For sensor failure cases, there are Motion Blur, Over-Exposure, and Under-Exposure common for RGB cameras, LiDAR-Jitter for LiDAR sensor and Event Low-resolution for event camera.
MCubeS is a multi-modal dataset with pairs of RGB, Near-Infrared (NIR), Degree of Linear Polarization (DoLP), and Angle of Linear Polarization (AoLP)

⋆ Corresponding author

of 20 category segmentation annotations. It has 302/96/102 image pairs for training/validation/testing at the size of 1224 $\times$ 1024.

Table 1: Per-class results on MCubeS [4] dataset. The training and validation is conducted with four modalities: Image, Aolp, Dolp, and Nir. (M-2: MiT-B2)

MCubeS	Method	Backbone	#Param(M)	Asph.	Concrete	Metal	R.M.	Fabr.	Glass	Plaster	Plastic	Rubber	Sand	Mean
IoU	[11]	M-2	58.73	84.71	45.23	54.10	74.80	32.18	54.34	0.69	28.54	28.72	67.80	-
	Ours	M-2	**24.73**	**88.86**	**50.79**	53.31	**74.98**	**38.55**	**55.27**	**0.86**	**34.21**	**31.93**	66.86	-
			Δ	**+4.15**	**+5.56**	-0.79	**+0.18**	**+6.37**	**+0.93**	**+0.17**	**+5.67**	**+3.21**	-0.94	-
	Method	Backbone	Param	Gravel	Ceramic	Cobb.	Brick	Grass	Wood	Leaf	Water	Human	Sky	Mean
	[11]	M-2	58.73	67.12	26.81	68.67	43.19	58.95	49.71	75.62	54.35	18.67	96.52	51.54
	Ours	M-2	**24.73**	66.80	**31.21**	**71.67**	**46.47**	55.26	48.72	75.33	**54.72**	17.91	**96.60**	**53.01**
			Δ	-0.32	**+4.40**	**+3.00**	**+3.28**	-3.69	-0.99	-0.29	**+0.37**	-0.76	**+0.08**	**+1.47**
MCubeS	Method	Backbone	#Param(M)	Asph.	Concrete	Metal	R.M.	Fabr.	Glass	Plaster	Plastic	Rubber	Sand	Mean
F1	[4]	M-2	58.73	91.72	62.29	70.21	85.59	48.69	70.41	1.37	44.40	44.62	80.81	-
	Ours	M-2	**24.73**	**94.10**	**67.36**	69.55	**85.70**	**55.64**	**71.19**	**1.70**	**50.98**	**48.40**	80.14	-
			Δ	**+2.38**	**+5.27**	-0.66	**+0.11**	**+6.95**	**+0.78**	**+0.33**	**+6.58**	**+3.78**	-0.67	-
	Method	Backbone	Param	Gravel	Ceramic	Cobb.	Brick	Grass	Wood	Leaf	Water	Human	Sky	Mean
	[11]	M-2	58.73	80.33	42.28	81.43	60.32	74.18	66.41	86.12	70.42	31.47	98.23	64.57
	Ours	M-2	**24.73**	80.09	**47.57**	**83.50**	**63.45**	71.18	65.52	85.93	**70.73**	30.37	**98.27**	**66.07**
			Δ	-0.24	**+5.29**	**+2.07**	**+3.13**	-3.00	-0.89	-0.19	**+0.31**	-1.10	**+0.04**	**+1.50**
MCubeS	Method	Backbone	#Param(M)	Asph.	Concrete	Metal	R.M.	Fabr.	Glass	Plaster	Plastic	Rubber	Sand	Mean
Acc	[4]	M-2	58.73	93.87	58.29	76.71	82.37	40.59	67.98	1.58	34.38	36.13	85.77	-
	Ours	M-2	**24.73**	**96.46**	**66.00**	74.85	81.95	**46.45**	67.73	1.33	**42.55**	**41.87**	**88.58**	-
			Δ	**+2.59**	**+7.71**	-1.86	-0.42	**+5.86**	-0.25	-0.25	**+8.17**	**+5.74**	**+2.81**	-
	Method	Backbone	Param	Gravel	Ceramic	Cobb.	Brick	Grass	Wood	Leaf	Water	Human	Sky	Mean
	[11]	M-2	58.73	75.86	32.37	77.30	65.59	72.73	58.58	90.13	56.63	23.29	98.30	61.42
	Ours	M-2	**24.73**	74.22	**37.99**	76.13	**68.08**	64.66	**62.65**	**90.34**	**56.73**	20.54	98.29	**62.87**
			Δ	-1.64	**+5.62**	-1.17	**+2.49**	-8.07	**+4.07**	**+0.21**	**+0.10**	-2.75	-0.01	**+1.45**

1.2 Implementation Details

We train MAGIC on 8 $\times$ A800 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 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 [11], while to 512 $\times$ 512 on MCubeS [5]. ImageNet-1K pre-trained weight is used as the pre-trained weight.

1.3 Metrics

To evaluate the performance of our MAGIC framework, three metrics are utilized, including Intersection over Union (IoU), F1 score, and Accuracy (Acc).
IoU, also known as the Jaccard index, measures the overlap between the predicted segmentation and the ground truth segmentation. It is calculated by di-

viding the intersection of the two segmentation maps by their union. IoU ranges from 0 to 1, with a higher value indicating better segmentation performance.
F1 score is a measure of the model's precision and recall. It is calculated by taking the harmonic mean of precision and recall, where precision is the ratio of true positives to the sum of true and false positives, and recall is the ratio of true positives to the sum of true positives and false negatives. F1 score ranges from 0 to 1, with a higher value indicating better segmentation performance.
Acc measures the percentage of correctly classified pixels in the segmentation map. It is calculated by dividing the number of correctly classified pixels by the total number of pixels in the segmentation map. Accuracy ranges from 0 to 1, with a higher value indicating better segmentation performance.

2 Additional Experiments

Due to the lack of space in the main paper, we provide more experimental results in this section.

2.1 Multi-modal Semantic Segmentation

In this subsection, we present a quantitative comparison of our MAGIC with the state-of-the-art CMNeXt [11] method on the MCubeS dataset using three semantic segmentation metrics, namely IoU, F1, and Acc. The experimental results are shown in Tab. 1, underscoring the marked edge our framework possesses over existing methods.

Intriguingly, despite our MAGIC model encompassing merely 42% of CMNeXt's parameters (with 24.73M against CMNeXt's 58.73M), it demonstrates superior performance across numerous categories in mIoU, such as Asphalt (**88.86%** vs. 84.71% $\rightarrow$ +**4.15%**$\uparrow$), Concrete (**50.79%** vs. 45.23% $\rightarrow$ +**5.56%**$\uparrow$), Fabric (**38.55%** vs. 32.18% $\rightarrow$ +**6.37%**$\uparrow$), and Plastic (**34.21%** vs. 28.54% $\rightarrow$ +**5.67%**$\uparrow$). Similarly, our MAGIC framework (**24.73M**) consistently outperforms CMNeXt in most of the categories in Acc, such as Asphalt (**96.46%** vs. 93.87% $\rightarrow$ +**2.59%**$\uparrow$), Concrete (**66.00%** vs. 58.29 $\rightarrow$ +**7.71%**$\uparrow$), Fabric (**46.45%** vs. 40.59 $\rightarrow$ +**5.86%**$\uparrow$), and Plastic (**42.55%** vs. 34.38% $\rightarrow$ +**8.17%**$\uparrow$). These results demonstrate the effectiveness of our plug-and-play modules over the prior Hub2Fuse, separate branch, and joint branch paradigms. Notably, for the mean performance on the three metrics, our MAGIC consistently outperforms the previous state-of-the-art method CMNeXt by +**1.47% mIoU**, +**1.50% mF1**, and +**1.45% mAcc**, respectively.

In Tab. 2, we showcase the performance of our MAGIC framework on the DELIVER dataset, which incorporates four modalities: RGB, Depth, Event, and LiDAR. Empirical results indicate that MAGIC consistently surpasses the state-of-the-art CMNeXt [11] with improvements of +1.33% in mIoU, +1.30% in mF1, and +0.92% in mAcc.

Further, MAGIC excels in several categories, notably Side Walk (**86.22%** compared to 82.27%, an increase of +**3.95%**) and Cars (**90.94%** compared

Table 2: Per-class results on DELIVER dataset. The training and validation is conducted with four modalities: RGB, Depth, Event, and LiDAR. (M-2: MiT-B2)

Metric	Method	Backbone	Param	Build.	Fence	Other	Pede.	Pole	RL	Road	Side W.	Veget.	Cars	Wall	T. S.	Sky
	[11]	M-2	58.73	89.41	43.12	0	76.51	75.13	85.91	98.18	82.27	88.97	84.98	69.39	70.57	99.43
	Ours	M-2	24.73	**89.66**	**49.27**	0	**76.54**	72.64	84.81	**98.40**	**86.22**	88.71	**90.94**	**70.86**	**72.88**	99.39
			Δ	**+0.25**	**+6.15**	0	**+0.03**	-2.49	-1.10	**+0.22**	**+3.95**	-0.26	**+5.96**	**+1.47**	**+2.31**	-0.04
IoU	Method	Backbone	Param	Ground	Bridge	Rail T.	G. R.	Traffic L.	Static	Dynamic	Water	Terr.	Two W.	Bus	Truck	Mean
	[11]	M-2	58.73	1.31	53.61	61.48	55.01	84.22	33.58	32.30	23.96	83.94	77.33	92.25	94.55	66.30
	Ours	M-2	24.73	**2.62**	**59.28**	59.76	**73.08**	82.76	**35.70**	30.23	**30.93**	**84.00**	76.22	84.58	91.99	**67.66**
			Δ	**+1.31**	**+5.67**	-1.72	**+18.07**	-1.46	**+2.12**	-2.07	**+6.97**	**+0.06**	-1.11	-7.67	-2.56	+1.33
Metric	Method	Backbone	Param	Build.	Fence	Other	Pede.	Pole	RL	Road	Side W.	Veget.	Cars	Wall	T. S.	Sky
	[11]	M-2	58.73	94.41	60.26	0	86.69	85.80	92.42	99.08	90.28	94.16	91.88	81.93	82.74	99.71
	Ours	M-2	24.73	**94.55**	**66.02**	0	**86.71**	84.15	91.78	**99.20**	**92.60**	94.02	**95.25**	**82.95**	**84.31**	99.69
			Δ	**+0.14**	**+5.76**	0	**+0.02**	-1.65	-0.64	**+0.12**	**+2.32**	-0.14	**+3.37**	**+1.02**	**+1.57**	-0.02
F1	[11]	M-2	58.73	2.59	69.80	76.14	70.98	91.43	50.28	48.83	38.66	91.27	87.22	95.97	97.20	75.19
	Ours	M-2	24.73	**5.11**	**74.43**	74.81	**84.44**	90.57	**52.61**	46.43	**47.25**	**91.30**	86.51	91.65	95.83	**76.49**
			Δ	**+2.52**	**+4.63**	-1.33	**+13.46**	-0.86	**+2.33**	-2.40	**+8.59**	**+0.03**	-0.71	-4.32	-1.37	+1.30
Metric	Method	Backbone	Param	Build.	Fence	Other	Pede.	Pole	RL	Road	Side W.	Veget.	Cars	Wall	T. S.	Sky
	[11]	M-2	58.73	98.24	57.18	0	87.58	85.25	89.70	98.95	95.36	94.19	98.65	87.98	83.33	99.75
	Ours	M-2	24.73	**98.42**	**62.14**	0	86.91	81.49	88.75	**99.22**	93.90	93.69	97.97	86.58	80.42	99.75
			Δ	**+0.18**	**+4.96**	0	-0.67	-3.76	-0.95	**+0.27**	-1.46	-0.50	-0.68	-1.40	-2.91	0.00
Acc	Method	Backbone	Param	Ground	Bridge	Rail T.	G. R.	Traffic L.	Static	Dynamic	Water	Terr.	Two W.	Bus	Truck	Mean
	[11]	M-2	58.73	2.00	61.91	75.28	56.60	88.71	35.32	50.35	24.05	93.65	86.86	96.13	97.12	73.77
	Our	M-2	24.73	**5.56**	**62.32**	73.23	**76.77**	**89.12**	**38.02**	49.93	**33.07**	93.39	83.79	**96.22**	96.64	**74.69**
			Δ	**+3.56**	**+0.41**	-2.05	**+20.17**	**+0.41**	**+2.70**	-0.42	**+9.02**	-0.26	-3.07	**+0.09**	-0.48	+0.92

to 84.98%, an increment of +**5.96%**). This underscores the robustness of the MAGIC framework, especially with the integration of the proposed MAM and ASM modules for multi-modal learning. It's important to mention that our evaluation on the DELIVER dataset also spanned Image, Aolp, Dolp, and Nir modalities, reinforcing the adaptability of our approach across varied modalities.

2.2 Adverse Weather and Sensor Failures

This section offers an in-depth assessment of our MAGIC framework in comparison to leading multi-modal fusion approaches, tested under a spectrum of adverse weather conditions such as cloudy, foggy, rainy, and sunny days. Additionally, we examine performance during partial sensor failure scenarios, including motion blur, over-exposure, under-exposure, LiDAR jitter, and reduced resolution, utilizing the DELIVER dataset.

Tab. 3 elucidates the comparative performance, underscoring the dominance of MAGIC, which employs SegFormer-B0, over established methods like HRFuser [9], TokenFusion [7], CMX [10], and CMNeXt [11]. Notably, when employing the RGB-Depth-LiDAR sensor combinations, our MAGIC surpasses CMNeXt [11] integrated with SegFormer-B0 by an impressive +**6.38%** mIoU average under challenging conditions. In a direct comparison to TokenFusion—limited to RGB-Depth sensors and equipped with 26.01M parameters—our streamlined CMNeXt model, boasting just **3.72M** parameters, yields a marked +**9.34%** mIoU improvement in mean performance during adverse settings.

Table 3: Results on adverse conditions of DELIVER. Sensor failures are MB: Motion Blur; OE: Over-Exposure; UE: Under-Exposure; LJ: LiDAR-Jitter; and EL: Event Low-resolution. The parameters (#Params) and GFLOPs are counted in 512 × 512.

Model-modality	#Param	Cloudy	Foggy	Night	Rainy	Sunny	MB	OE	UE	LJ	EL	Mean
HRFuser [9]-RGB	29.89	49.26	48.64	42.57	50.61	50.47	48.33	35.13	26.86	49.06	49.88	47.95
SegFormer [8]-RGB	25.79	59.99	57.30	50.45	58.69	60.21	57.28	56.64	37.44	57.17	59.12	57.20
TokenFusion [7]-RGB-D	26.01	50.92	52.02	43.37	50.70	52.21	49.22	46.22	36.39	49.58	49.17	49.86
CMX [10]-RGB-D	66.57	63.70	62.77	60.74	62.37	63.14	59.50	60.14	55.84	62.65	63.26	62.66
HRFuser [9]-RGB-D	30.46	54.80	51.48	49.51	51.55	52.12	50.92	41.51	44.00	54.10	52.52	51.88
HRFuser [9]-RGB-D-E	31.04	54.04	50.83	50.88	51.13	52.61	49.32	41.75	47.89	54.65	52.33	51.83
HRFuser [9]-RGB-D-E-L	31.61	56.20	52.39	49.85	52.53	54.02	49.44	46.31	46.92	53.94	52.72	52.97
CMNeXt [11] w/ M-0 RGB-D-L	58.69	56.34	54.53	51.19	51.64	54.90	49.63	54.50	48.08	56.45	50.89	52.82
CMNeXt [11] w/ M-0 RGB-D-E	58.72	56.61	52.83	51.33	53.97	55.53	50.63	54.69	48.99	56.28	52.54	53.34
CMNeXt [11] w/ M-0 RGB-D-E-L	58.73	60.06	56.16	54.03	54.82	58.29	53.70	57.04	51.98	58.54	55.87	56.01
MAGIC w/ M-0 RGB-D-L	3.72	60.96	62.03	58.42	57.24	61.18	57.16	59.27	57.43	61.12	57.15	59.20
w.r.t CMNeXt [11] w/ M-0 RGB-D-L		**+4.62**	**+7.50**	**+7.23**	**+5.60**	**+6.28**	**+7.53**	**+4.77**	**+9.35**	**+4.67**	**+6.26**	**+6.38**
MAGIC w/ M-0 RGB-D-E	3.72	63.67	62.21	61.66	59.95	64.43	60.21	61.31	60.91	62.59	61.08	61.80
w.r.t CMNeXt [11] w/ M-0 RGB-D-E		**+7.06**	**+9.38**	**+10.33**	**+5.98**	**+8.90**	**+9.58**	**+6.62**	**+11.92**	**+6.31**	**+8.54**	**+8.46**
MAGIC w/ M-0 RGB-D-E-L	3.72	65.90	62.52	61.17	63.87	61.17	63.87	62.44	59.62	63.42	61.70	62.57
w.r.t CMNeXt [11] w/ M-0 RGB-D-E-L		**+5.84**	**+6.36**	**+7.14**	**+9.05**	**+2.88**	**+10.17**	**+5.40**	**+7.64**	**+4.88**	**+5.83**	**+6.56**
CMNeXt [11] w/ M-2 RGB-D-L	58.69	67.21	62.79	61.64	62.95	65.26	61.00	64.64	58.71	64.32	63.35	63.58
CMNeXt [11] w/ M-2 RGB-D-E	58.72	68.28	63.28	62.64	63.01	66.06	62.58	64.44	58.73	65.37	65.80	64.02
CMNeXt [11] w/ M-2 RGB-D-E-L	58.73	68.70	65.67	62.46	67.50	66.57	62.91	64.59	60.00	65.92	65.48	64.98
MAGIC w/ M-2 RGB-D-L	24.73	68.64	66.59	67.17	67.02	67.58	63.93	65.68	65.58	67.46	66.43	66.61
w.r.t CMNeXt [11] w/ M-2 RGB-D-L		**+1.43**	**+3.80**	**+5.53**	**+5.38**	**+2.32**	**+2.93**	**+1.04**	**+6.87**	**+3.14**	**+3.08**	**+3.03**
MAGIC w/ M-2 RGB-D-E	24.73	67.15	65.41	64.74	66.09	66.66	63.83	64.77	63.59	66.24	63.68	65.22
w.r.t CMNeXt [11] w/ M-2 RGB-D-E		-1.13	**+2.13**	**+2.10**	**+3.08**	**+0.60**	**+1.25**	**+0.33**	**+4.86**	**+0.87**	-2.12	**+1.20**
MAGIC w/ M-2 RGB-D-E-L	24.73	68.89	67.23	66.54	67.06	66.62	65.10	64.14	63.51	67.14	67.36	66.36
w.r.t CMNeXt [11] w/ M-2 RGB-D-E-L		**+0.19**	**+1.56**	**+4.08**	-0.44	**+0.05**	**+2.19**	-0.45	**+3.51**	**+1.22**	**+1.88**	**+1.38**

Further deepening the analysis, MAGIC consistently demonstrates superior performance over CMNeXt [11] across the majority of sensor malfunction scenarios detailed in Tab. 3. Specifically, with the integration of SegFormer-B2, we witness performance boosts of **+3.51%, +1.22%, and +1.88%** mIoU during under-exposure, LiDAR-jitter, and event low-resolution situations respectively.

Remarkably, for under-exposure scenarios, MAGIC overshadows the RGB baseline with an impressive **+26.07%** mIoU, attesting to its resilience in strenuous circumstances. The empirical results accentuate the invaluable contributions of our MAM and ASM in refining MAGIC's performance relative to its predecessors.

2.3 Comparison with State-of-the-Art Methods

Results on DELIVER: Tab. 4 presents a comprehensive comparison of our MAGIC framework with other state-of-the-art methods for fusing RGB with Depth, Event, and LiDAR modalities. The results demonstrate that our MAGIC

Table 4: Results of multi-modal semantic segmentation on DELIVER.

Method	Modal	Backbone	mIoU(%)	Method	Modal	Backbone	mIoU(%)
HRFuser [9]	R	HRFuser-T	47.95	HRFuser [9]	R+L	HRFuser-T	43.13
		MiT-B0	52.10	TF [7]	R+L	MiT-B2	53.01
SegFormer [8]	R	MiT-B1		CMX [10]	R+L	MiT-B2	56.37
		MiT-B2	57.20	CMNeXt [11]	R+L	MiT-B2	58.04
HRFuser [9]	R+D	HRFuser-T	51.88	Ours	R+L	MiT-B2	57.75
TokenFusion [7]	R+D	MiT-B2	60.25	HRFuser [9]	R+D+E	HRFuser-T	51.83
CMX [10]	R+D	MiT-B2	62.67	CMNeXt [11]	R+D+E	MiT-B2	64.44
CMNeXt [11]	R+D	MiT-B2	63.58	Ours	R+D+E	MiT-B2	**66.24** +1.80↑
Ours	R+D	MiT-B2	**66.89** +3.31↑	HRFuser [9]	R+D+L	HRFuser-T	52.72
HRFuser [9]	R+E	HRFuser-T	42.22	CMNeXt [11]	R+D+L	MiT-B2	65.50
TF [7]	R+E	MiT-B2	45.63	Ours	R+D+L	MiT-B2	**67.63** +2.13↑
CMX [10]	R+E	MiT-B2	56.62	HRFuser [9]	R+D+E+L	HRFuser-T	52.97
CMNeXt [11]	R+E	MiT-B2	57.48	CMNeXt [11]	R+D+E+L	MiT-B2	66.30
Ours	R+E	MiT-B2	**58.48** +1.00↑	Ours	R+D+E+L	MiT-B2	**67.66** +1.36↑

framework outperforms other methods in terms of multi-modal semantic segmentation performance. Specifically, our dual modality MAGIC framework achieves superior performance compared to HRFuser [9], TokenFusion [7], CMX [10], and CMNeXt [11] in most fusion scenarios.

Notably, when training with both RGB and Depth data, our MAGIC outperforms the previous state-of-the-art method CMNeXt by +3.31% mIoU. Moreover, when training with RGB, Depth, and LiDAR data, our MAGIC achieves 67.63% mIoU, which is +2.13% mIoU higher than the performance of CMNeXt. These results demonstrate the effectiveness of our proposed MAGIC framework for multi-modal semantic segmentation.

Tab. 5 presents a comprehensive comparison of our MAGIC framework with other state-of-the-art methods for fusing RGB with Image, Aolp, Dolp, and NIR modalities on the MCubeS dataset. The results demonstrate that our MAGIC framework outperforms other methods in terms of multi-modal segmentation performance. Specifically, our dual modality MAGIC framework achieves superior performance compared to DRConv [1], DDF [12], TransFuser [6], MMTM [3], FuseNet [2], MCubeSNet [4], and CMNeXt [11]. Notably, when training with Image, Aolp, and Dolp data, our MAGIC achieves 52.83% mIoU, which is +3.35% mIoU higher than the performance of CMNeXt.

Overall, the proposed MAGIC framework represents a significant advancement in the field of multi-modal semantic segmentation, offering a powerful and effective way to fuse different modalities for more accurate and efficient image segmentation.

2.4 Modality-agnostic Segmentation

In this subsection, we introduce the modality-agnostic segmentation, which differs from the approach proposed in [11], where the arbitrary modality inputs cannot be without the RGB data. In our paper, we utilize arbitrary inputs without relying on each of the modalities. To verify the robustness of our proposed

Table 5: Results of multi-modal segmentation on MCubeS.

Method	Modal	mIoU
DRConv [1]	I-A-D-N	34.63
DDF [12]	I-A-D-N	36.16
TransFuser [6]	I-A-D-N	37.66
MMTM [3]	I-A-D-N	39.71
FuseNet [2]	I-A-D-N	40.58
MCubeSNet [4]	I	33.70
MCubeSNet [4]	I-A	39.10
MCubeSNet [4]	I-A-D	42.00
MCubeSNet [4]	I-A-D	42.86
CMNeXt [11] (MiT-B2)	I	48.16
CMNeXt [11] (MiT-B2)	I-A	48.82
CMNeXt [11] (MiT-B2)	I-A-D	49.48
CMNeXt [11] (MiT-B2)	I-A-D-N	51.54
MAGIC (MiT-B2)	I	-
MAGIC (MiT-B2)	I-A	-
MAGIC (MiT-B2)	I-A-D	**52.83**
w.r.t CMNeXt	-	**+3.35**
MAGIC (MiT-B2)	I-A-D-N	**53.01**
w.r.t CMNeXt	-	**+1.47**

Table 6: Results of MAGIC validation with 2 modalities on DELIVER. (M-0: MiT-B0; M-2: MiT-B2)

Train	Method	Backbone	#Param(M)	MAGIC Validation							Mean	Δ
				R	D	E	L	R+D	R+E	R+L		
R+D	CMNeXt	M-2	58.69	1.60	1.44	-	-	63.58	-	-	22.81	-
	Ours	M-0	3.72	30.47	56.44	-	-	63.46	-	-	**38.87**	**+16.06**
		M-2	24.73	37.26	59.02	-	-	66.89	-	-	**54.39**	**+31.58**
R+E	CMNeXt	M-2	58.69	4.82	-	3.45	-	-	57.48	-	21.92	-
	Ours	M-0	3.72	52.63	-	11.28	-	-	52.69	-	**38.87**	**+16.95**
		M-2	24.73	58.00	-	14.81	-	-	58.48	-	**43.76**	**+21.84**
R+L	CMNeXt	M-2	58.69	2.10	-	-	2.56	-	-	58.04	20.90	-
	Ours	M-0	3.72	51.55	-	-	15.75	-	-	53.01	**40.10**	**+19.20**
		M-2	24.73	57.13	-	-	19.46	-	-	57.75	**44.78**	**+23.88**

Table 7: Results of MAGIC validation with 2 modalities on MCubeS [4]. (M-0: MiT-B0; M-2: MiT-B2)

Train	Method	Backbone	#Param(M)	MAGIC Validation							Mean	*Δ*
				I	A	D	N	I+A	I+D	I+N		
I+A	CMNeXt	M-2	58.69	3.99	1.74	-	-	29.31	-	-	11.68	-
	Ours	M-0	3.72	32.92	23.02	-	-	42.71	-	-	**32.88**	**+21.20**
		M-2	24.73	51.45	0.27	-	-	51.45	-	-	**34.39**	**+22.71**
I+D	CMNeXt	M-2	58.69	2.26	-	0.71	-	-	33.02	-	12.00	-
	Ours	M-0	3.72	45.98	-	20.71	-	-	45.89	-	**37.53**	**+25.53**
		M-2	24.73	49.93	-	0.06	-	-	49.96	-	**33.32**	**+21.32**
I+N	CMNeXt	M-2	58.69	2.14	-	-	1.53	-	-	33.39	12.35	-
	Ours	M-0	3.72	36.81	-	-	8.38	-	-	44.60	**29.93**	**+17.58**
		M-2	24.73	51.20	-	-	3.03	-	-	51.69	**35.31**	**+22.96**

method with arbitrary modality inputs, we apply the MAGIC framework on both the DELIVER and MCubeS [5] datasets.

Tab. 6 presents the results of our MAGIC framework trained with two selected modalities, namely RGB+Depth, RGB+Event, and RGB+LiDAR, and validated with arbitrary modality combinations[3]. Our MAGIC significantly outperforms CMNeXt [11] in all validation scenarios, achieving a performance gain of +16.06% and +31.58% mIoU compared to CMNeXt [11] with SegFromer-B0 and -B2 backbone on the DELIVER [11] dataset with RGB+Depth scenario, respectively. On the MCubeS [5] dataset, our MAGIC significantly outperforms CMNeXt [11] in all validation scenarios, achieving a performance gain of +21.20% and +22.71% mIoU compared to CMNeXt [11] with SegFromer-B0 and -B2 backbone on the Image+Aolp scenario, respectively.

Notably, in the RGB data absence validation scenarios, our MAGIC with SegFormer-B2 demonstrates a significant performance gain, such as Depth only (**59.02** vs. 1.44 $\rightarrow$ **+57.58**$\uparrow$). Moreover, our MAGIC with SegFormer-B2 has only 42% of the parameters of CMNeXt. Furthermore, our MAGIC with SegFormer-B0 surpasses CMNeXt [11] by a large margin with only **0.06%** parameters. These results demonstrate the effectiveness of our proposed MAM and ASM modules as powerful plug-and-play modules for multi-modal visual learning, especially for arbitrary modality input scenarios.

2.5 Extension to diverse models

We have implemented MAGIC across different backbones, including the LiteSeg framework with MobileNet and FPN with PVTv2-b0, as in Tab. 11 and Tab. 12. The results consistently demonstrate the robustness and superiority of MAGIC across different backbones, from CNNs to ViTs.

[3] Since CMNeXt cannot be implemented without the RGB input, we compare the RGB+X settings for dual modality semantic segmentation

Table 8: Ablation study of the selection of the salient features in AFLM on MCubeS [5] w/ MiT-b0.

Salient Features	○ ○ ○ ○	⋆ ○ ○ ○	○ ○ ○ ⋆	⋆ ○ ○ ⋆	○ ○ ⋆ ⋆	○ ⋆ ○ ⋆	⋆ ○ ⋆ ○	⋆ ⋆ ○ ○	○ ⋆ ⋆ ○
DELIVER	59.26	62.13	59.93	**62.19**	61.83	59.31	62.06	59.93	59.78
MCubeS	○ ○ ○ ○	⋆ ○ ○ ○	○ ○ ○ ⋆	⋆ ○ ○ ⋆	○ ○ ⋆ ⋆	○ ⋆ ○ ⋆	⋆ ○ ⋆ ○	⋆ ⋆ ○ ○	○ ⋆ ⋆ ○
mIoU (%)	42.43	46.39	44.11	**46.50**	46.47	46.01	43.01	44.09	44.43

Table 9: Ablation study of components in MAM with MiT-B0 on MCubeS.

MAM	w/o Residual Block	w/o Parallel Pooling	w/o MLP	All
mIoU	46.00	45.91	47.14	**47.58**

Pooling Size	(1,3,5)	(3,5,7)	(3,7,11)	(5,7,11)	(7,11,21)
mIoU	45.65	45.49	**47.58**	45.23	44.99

Table 10: Ablation study of λ and β on MCubeS [5].

λ	0.2	0.1	0.06	0.05	0.04	0.02
mIoU	51.88	52.07	52.17	**52.24**	51.62	51.26
β (w/ λ=0.05)	0.5	1	2	4	6	10
mIoU	52.39	52.37	**53.01**	51.97	52.19	52.16

Table 11: Multi-modal segmentation comparison on DELIVER.

	Backbone	Modal (DELIVER)	mIoU	Δ
CMNeXt	FPN + PVT-v2-B0	RGB+Event	51.50	-
MAGIC			58.78	+7.28
CMNeXt	FPN + PVT-v2-B0	RGB+Depth	55.25	-
MAGIC			61.22	+5.97
CMNeXt	FPN + PVT-v2-B0	R-D-E-L	61.52	-
MAGIC			66.33	+4.81
Baseline	LiteSeg + MobileNet	RGB	29.65	-
CMNeXt	LiteSeg + MobileNet	R-D-E-L	56.54	-
MAGIC			61.58	+5.04

Table 12: Multi-modal segmentation comparison on MCubeS.

	Backbone	Modal (MCubeS)	mIoU	Δ
CMNeXt	LiteSeg + MobileNet	Image+Nir	33.39	-
MAGIC			51.20	+17.81
CMNeXt	LiteSeg + MobileNet	Image+AoLP	29.31	-
MAGIC			51.44	+22.13
CMNeXt	LiteSeg + MobileNet	Image+DoLP	33.02	-
MAGIC			49.96	+16.94

Qualitative Results This subsubsection presents a qualitative comparison of the semantic segmentation results obtained using our MAGIC and CMNeXt [11]

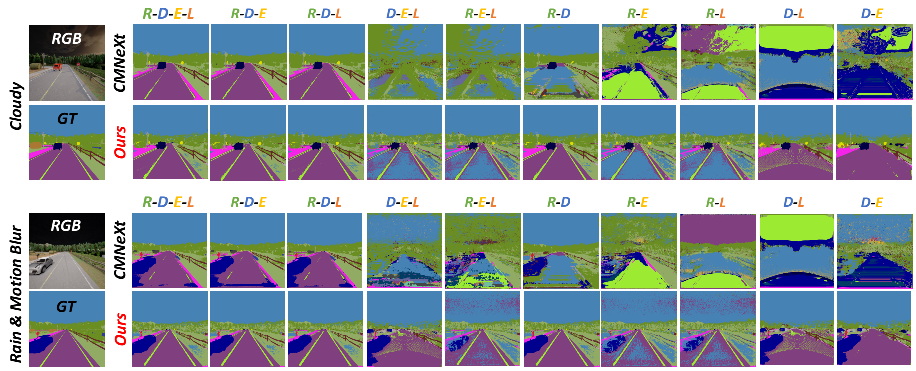

Fig. 1: Visualization of arbitrary inputs using {**R**GB, **D**epth, **E**vent, **Li**DAR} on DELIVER. CMNeXt: results of CMNeXt [11]; Ours: results of our MAGIC, on normal conditions, *i.e.*, cloudy and rain with motion blur.

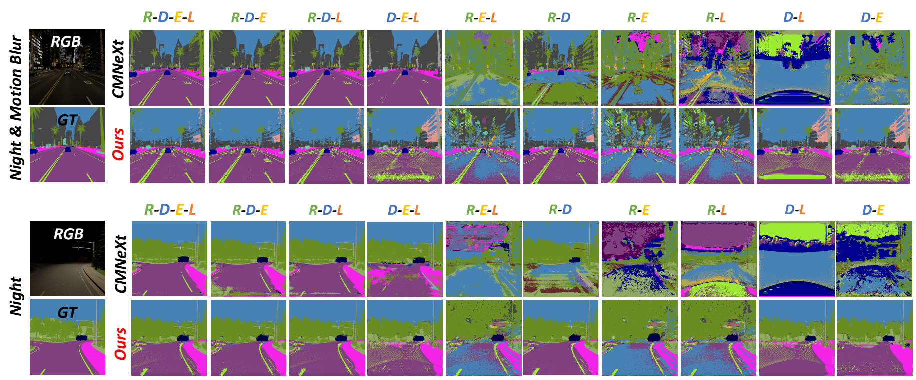

Fig. 2: (a) & (b): Visualization of arbitrary inputs using {**R**GB, **D**epth, **E**vent, **Li**DAR} on DELIVER. (a): results of CMNeXt [11]; (b): results of our MAGIC, on challenging conditions, *i.e.*, night with motion blur and night.

in various autonomous driving scenes, including normal, challenging, and extreme scenarios.

In Fig.1, we present more visual comparisons of the semantic segmentation results obtained using our MAGIC and CMNeXt [11] in normal autonomous driving scenes, such as rainy weather and motion blur. The results demonstrate that our MAGIC consistently performs well with arbitrary inputs, whereas CMNeXt is fragile in most scenarios.

In Fig.2, we present more visual comparisons of the semantic segmentation results obtained using our MAGIC and CMNeXt [11] in challenging autonomous driving scenes, such as night driving with motion blur. The results demonstrate that our MAGIC consistently performs well with arbitrary inputs, whereas CMNeXt is fragile in most scenarios.

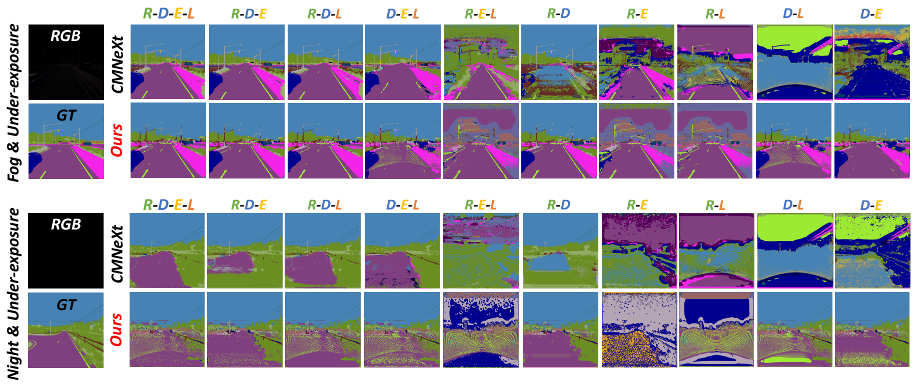

Fig. 3: (a) & (b): Visualization of arbitrary inputs using {**R**GB, **D**epth, **E**vent, **Li**DAR} on DELIVER. CMNeXt: results of CMNeXt [11]; Ours: results of our MAGIC, on extreme conditions, *i.e.*, night with under-exposure and fog with under-exposure.

In Fig.3, we present more visual comparisons of the semantic segmentation results obtained using our MAGIC and CMNeXt [11] in extreme autonomous driving scenes, such as driving at night with under-exposure lighting condition. 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 sensing data, which further enhances the robustness of full scene segmentation under varying lighting and weather conditions, such as cloudy, rain, and motion blur. These qualitative results further demonstrate the effectiveness and robustness of our proposed MAGIC framework.

3 Additional Ablation Study

Ablation Study of MAM Components As indicated in Tab. 9, we conduct an ablation study of the components in the proposed MAM. Removing any of the components, namely the residual block, parallel pooling, and MLP, leads to a drop in performance. Therefore, all of the components play a positive and crucial role in our MAM.

Ablation of Pooling Size We ablate the pooling size in parallel pooling within MAM on the MCubeS dataset, as presented in Tab. 9. Our results demonstrate that the pooling size of (3,7,11) achieves the best mIoU.

Ablations of the Hyper-parameters λ and β We now investigate the impact of hyper-parameters λ and β, which represent the weights for loss functions $\mathcal{L}_A$ and $\mathcal{L}_C$, respectively. Tab. 10 presents the experimental results for varying values of λ and β.

Algorithm 1: Framework of our proposed MAGIC.

Input: RGB images $R \in \mathbb{R}^{h \times w \times 3}$, depth maps $D \in \mathbb{R}^{h \times w \times C^D}$, LiDAR point cloud $L \in \mathbb{R}^{h \times w \times C^L}$, event streams $E \in \mathbb{R}^{h \times w \times C^E}$, and the corresponding groudn truth y with K categories

Initialize backbone model with Imagenet1K pre-trained weights and randomly initialize our MAM and ASM;

for *each epoch* **do**

  for *each iteration* **do**

    1. Sample a mini-batchr, d, l, e, and y with K categories;
    2. Get features with backbone model: $\{f_r, f_d, f_l, f_e\} = F(\{r, d, l, e\})$;
    3. Pass $\{f_r, f_d, f_l, f_e\}$ to the MAM to get semantic features f_{sa};
    4. Pass f_{sa} to the seghead to get the predictions P_m to be supervised from y: $\mathcal{L}_M = -\sum_0^{K-1} y \cdot log(P_m)$;
    5. Cross-modal semantic similarity ranking with $\{f_r, f_d, f_l, f_e\}$ with the semantic features f_{sa} obtained from MAM, thereby deriving a similarity ranking and find the salient features f_{sa} and the remaining features f_{rm}: $f_{sa}, f_{rm} = Rank(Cos(\{f_r, f_d, f_l, f_e\}, f_{sa}))$;
    6. f_{sa} are then passed to another MAM for generating predictions P_s with the seghead;
    7. y smoothness ;
    8. Arbitrary-modal learning loss: $\mathcal{L}_S = -\sum_0^{K-1} y \cdot log(P_s)$;
    9. Semantic consistency training between the remaining features f_{rm}: $\mathcal{L}_C = \sum_0^C (c_1 log \frac{c_1}{\frac{1}{2}(c_1+c_2)} + c_2 log \frac{c_2}{\frac{1}{2}(c_1+c_2)})$;
    10. Total loss function: $\mathcal{L} = \mathcal{L}_M + \lambda \mathcal{L}_A + \beta \mathcal{L}_C$;
    11. Loss backwards and update parameters of the backbone model and our MAM and ASM.

  end

end

Visualization of Semantic and Salient Features We visualize the RGB image features, semantic features extracted by MAM, and the salient features extracted by ASM in Fig. 4 (e). Obviously, the semantic and the salient features capture better scene details compared with the RGB features, indicating that our proposed MAM and ASM successfully take advantages of the multi-modal input data.

4 Algorithm

Algorithm 1 describes the training procedure for our multi-modal fusion and segmentation model, MAGIC, that processes RGB images, depth maps, LiDAR point clouds, and event streams to predict pixel-wise semantic categories.

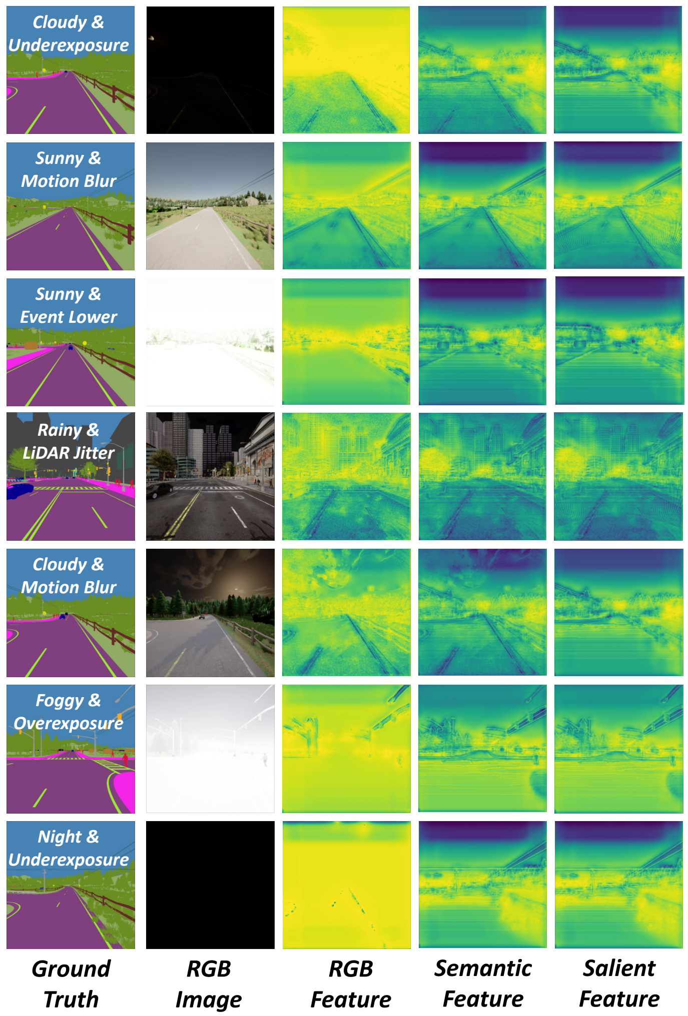

Fig. 4: Visualization of Semantic and Salient Features.

References

1. Chen, J., Wang, X., Guo, Z., Zhang, X., Sun, J.: Dynamic region-aware convolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8064–8073 (2021)
2. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13. pp. 213–228. Springer (2017)
3. Joze, H.R.V., Shaban, A., Iuzzolino, M.L., Koishida, K.: Mmtm: Multimodal transfer module for cnn fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13289–13299 (2020)
4. 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)
5. 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)
6. Prakash, A., Chitta, K., Geiger, A.: Multi-modal fusion transformer for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7077–7087 (2021)
7. 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)
8. 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)
9. Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J.: Hrformer: High-resolution vision transformer for dense predict. Advances in Neural Information Processing Systems **34**, 7281–7293 (2021)
10. 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)
11. 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)
12. Zhou, J., Jampani, V., Pi, Z., Liu, Q., Yang, M.H.: Decoupled dynamic filter networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6647–6656 (2021)