Learning Modality-agnostic Representation for Semantic Segmentation from Any Modalities —Supplementary Material—

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

[1] Hong Kong University of Science and Technology, Guangzhou, China
zhengxu128@gmail.com, yuanhuiyilv@hkust-gz.edu.cn
[2] Hong Kong University of Science and Technology, Hong Kong, China
linwang@ust.hk
https://vlislab22.github.io/Any2Seg/

Abstract. Due to the limited space in the main paper, this supplementary material provides an expansive elucidation of the proposed method and additional experimental results. Sec. 1 offers more details in implementing our AnySeg framework. Sec. 2 presents more experimental results, emphasizing both quantitative and qualitative assessments, and Sec. 3 shows more ablation results.

1 Implementation Details.

1.1 Datasets.

The **DELIVER** dataset [2], as introduced by Zhang *et al.*, represents a significant multi-modal segmentation dataset leveraging the CARLA simulator to include a variety of data types such as Depth, LiDAR, Views, Event, and RGB data. This dataset is notable for its dual-case offerings, which encompass diverse environmental conditions—cloudy, foggy, night, rainy, and sunny weather—and five specific scenarios of partial sensor failures. These environmental conditions pose perceptual challenges through variations in sunlight positioning and intensity, atmospheric diffusion, precipitation effects, and scene shading. The sensor failure scenarios comprehensively simulate malfunctions common to RGB cameras (Motion Blur, Over-Exposure, Under-Exposure), LiDAR jitter, and reduced resolution in event cameras, thereby providing a robust platform for testing and developing perception algorithms under a wide range of conditions. Concurrently, **MCubeS** [1] offers a multi-modal dataset designed for segmentation tasks across 20 categories, featuring pairs of RGB, Near-Infrared (NIR), Degree of Linear Polarization (DoLP), and Angle of Linear Polarization (AoLP) images. The dataset is divided into training, validation, and testing subsets, with 302, 96, and 102 image pairs respectively. This arrangement enables thorough evaluation and advancement of segmentation models, incorporating polarization data alongside traditional imaging modalities.

* Corresponding author

1.2 Implementation Details.

Our Any2Seg is trained on 8 NVIDIA GPUs, starting with a learning rate of 6 e^{-5}, adjusted by a poly strategy (power 0.9) across 200 epochs, including an initial 10-epoch warm-up at 0.1 times the learning rate. We use the AdamW optimizer (epsilon $1e^{-8}$, weight decay $1e^{-2}$) with a batch size of 1 per GPU. Data augmentation includes random resizing (0.5-2.0 ratio), horizontal flipping, color jitter, gaussian blur, and cropping to 1024 $\times$ 1024 on [2] and 512 $\times$ 512 on [1].

1.3 Metrics.

In the evaluation of the performance of the proposed MAGIC framework, the assessment is anchored on three pivotal metrics: Intersection over Union (IoU), F1 score, and Accuracy (Acc), each providing unique insights into the model's segmentation capability.

Intersection over Union (IoU) The IoU metric, also recognized as the Jaccard index, serves as a quantitative measure of the extent of overlap between the predicted and the ground truth segmentation maps. It is computed as the quotient of the intersection and the union of the predicted and ground truth segmentation areas. The IoU metric is normalized to range between 0 and 1, where a value closer to 1 denotes superior segmentation accuracy.

F1 Score The F1 score, a harmonic mean of precision and recall, offers a balanced measure of the model's precision (the proportion of true positive results in all positive predictions) and recall (the proportion of true positive results among all actual positives). This metric is designed to provide a single measure to assess the precision-recall trade-off, with its value also ranging from 0 to 1, where higher values indicate more effective segmentation performance.

Accuracy (Acc) Accuracy, expressed as a proportion, measures the fraction of pixels in the segmentation map that are correctly classified. This metric is calculated by dividing the tally of correctly classified pixels by the total pixel count within the map. Like IoU and F1 score, accuracy values span from 0 to 1, with higher values reflecting enhanced segmentation accuracy.

These metrics collectively facilitate a comprehensive evaluation of the MAGIC framework's performance, enabling a nuanced analysis of its segmentation efficacy across diverse conditions.

2 Experimental Results

Tab. 1 shows the per-class results on DELIVER dataset, the training and validation is conducted with all four modalities. Tab. 2 gives the qualitative results with three metrics in MISS validation. Fig. 1 presents a qualitative analysis comparing the performance of our method with the MISS validation criterion, employing solely depth data for inference. Fig. 2 depicts a qualitative comparison, utilizing the MISS validation, with the inference phase incorporating both

depth and event data. Fig. 3 offers a qualitative comparison, following the MISS validation framework, with inference leveraging depth, event, and LiDAR data. Fig. 4 provides a visualization comparison under the MISS validation scheme, utilizing RGB and depth data for inference.

Table 1: Per-class results on DELIVER dataset. The training and validation is conducted with four modalities: RGB, Depth, Event, and LiDAR. (Seg-B2: MiT-B2)

			Param	Build.	Fence	Other	Pede.	Pole	RL	Road	Side W.	Veget.	Cars	Wall	T. S.	Sky
	[2]	Seg-B2	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	Seg-B2	24.73	89.59	45.34	0	78.49	75.91	85.87	98.33	84.51	88.95	91.56	58.30	71.97	99.45
		Δ		**+0.18**	**+2.22**	0	**+1.98**	**+0.78**	-0.04	-0.15	**+2.24**	-0.02	**+6.58**	-11.09	**+1.40**	**+0.02**
IoU	Method	Backbone	Param	Ground	Bridge	Rail T.	G. R.	Traffic L.	Static	Dynamic	Water	Terr.	Two W.	Bus	Truck	Mean
	[2]	Seg-B2	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	Seg-B2	24.73	2.30	59.80	66.59	63.50	85.04	37.36	33.41	46.22	82.65	78.08	91.61	91.30	68.25
		Δ		**+0.99**	**+6.19**	**+5.11**	**+8.49**	**+0.82**	**+3.78**	**+1.11**	**+22.26**	-1.29	**+0.75**	-0.64	-3.25	**+1.95**
Metric	Method	Backbone	Param	Build.	Fence	Other	Pede.	Pole	RL	Road	Side W.	Veget.	Cars	Wall	T. S.	Sky
	[2]	Seg-B2	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	Seg-B2	24.73	94.51	62.39	0	87.95	86.30	92.40	99.16	91.60	94.15	95.60	73.66	83.70	99.72
		Δ		**+0.10**	**+2.13**	0	**+1.26**	**+0.50**	-0.02	**+0.08**	**+1.32**	-0.01	**+3.72**	-8.27	**+0.96**	**+0.01**
F1	Method	Backbone	Param	Build.	Fence	Other	Pede.	Pole	RL	Road	Side W.	Veget.	Cars	Wall	T. S.	Sky
	[2]	Seg-B2	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	Seg-B2	24.73	4.51	74.85	79.95	77.67	91.92	54.39	50.08	63.22	90.50	87.69	95.62	95.45	77.08
		Δ		**+1.92**	**+5.05**	**+3.81**	**+6.69**	**+0.49**	**+4.11**	**+1.25**	**+24.56**	-0.77	**+0.47**	-0.35	-1.75	**+1.89**
Metric	Method	Backbone	Param	Build.	Fence	Other	Pede.	Pole	RL	Road	Side W.	Veget.	Cars	Wall	T. S.	Sky
	[2]	Seg-B2	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	Seg-B2	24.73	98.45	61.19	0	89.77	85.52	90.05	98.95	95.35	93.94	97.32	78.65	83.21	99.75
		Δ		**+0.21**	**+4.01**	0	**+2.19**	**+0.27**	**+0.35**	0	-0.01	-0.25	-1.33	-9.33	-0.12	0
Acc	Method	Backbone	Param	Ground	Bridge	Rail T.	G. R.	Traffic L.	Static	Dynamic	Water	Terr.	Two W.	Bus	Truck	Mean
	[2]	Seg-B2	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	Seg-B2	24.73	6.08	63.18	73.55	65.63	89.22	40.67	53.38	52.08	93.36	85.79	95.22	98.73	75.56
		Δ		**+4.08**	**+1.27**	-1.73	**+9.03**	**+0.51**	**+5.35**	**+3.03**	**+28.03**	-0.29	-1.07	-0.91	**+1.61**	**+1.79**

3 Ablation Study

Fig. 5 shows more visualization of RGB, depth, and our obtained modality-agnostic features. Fig. 6 presents more t-SNE visualization to ablate the effectiveness of our proposed inter- and intra-modal knowledge distillation.

Table 2: Results of system-level MISS evaluation. All methods are trained with four modalities, and the metric is mIoU for all numbers.

M.	Modality-incomplete Validation on DELIVER [2] (**mIoU**)															Mean	Δ
	R	D	E	L	RD	RE	RL	DE	DL	EL	RDE	RDL	REL	DEL	RDEL		
[2]	3.76	0.81	1.00	**0.72**	50.33	13.23	18.22	21.48	3.83	**2.86**	66.24	66.43	15.75	46.29	66.30	25.25	-
Ours	**39.02**	**60.11**	**2.07**	0.31	**68.21**	**39.11**	**39.04**	**60.92**	**60.15**	1.99	**68.24**	**68.22**	**39.06**	**60.95**	**68.25**	**45.04**	**+19.79**
M.	Modality-incomplete Validation on DELIVER [2] (**Acc**)															Mean	Δ
	R	D	E	L	RD	RE	RL	DE	DL	EL	RDE	RDL	REL	DEL	RDEL		
[2]	53.43	67.66	5.12	**4.07**	72.98	53.17	52.84	67.65	67.58	4.47	72.93	72.90	52.53	67.54	72.84	52.51	-
Ours	**57.82**	**69.06**	**7.06**	3.90	**75.57**	**57.84**	**57.66**	**69.90**	**69.05**	**7.06**	**75.59**	**75.54**	**57.67**	**69.87**	**75.56**	**55.28**	**+2.77**
M.	Modality-incomplete Validation on DELIVER [2] (**F1**)															Mean	Δ
	R	D	E	L	RD	RE	RL	DE	DL	EL	RDE	RDL	REL	DEL	RDEL		
[2]	45.02	67.36	3.11	**2.37**	74.36	45.55	45.75	67.47	67.38	2.65	74.37	74.36	46.02	67.44	74.37	50.51	-
Ours	**50.35**	**69.98**	**3.80**	0.60	**77.06**	**50.58**	**50.50**	**71.26**	**70.01**	**3.66**	**77.08**	**77.06**	**50.67**	**71.28**	**77.08**	**53.40**	**+2.89**

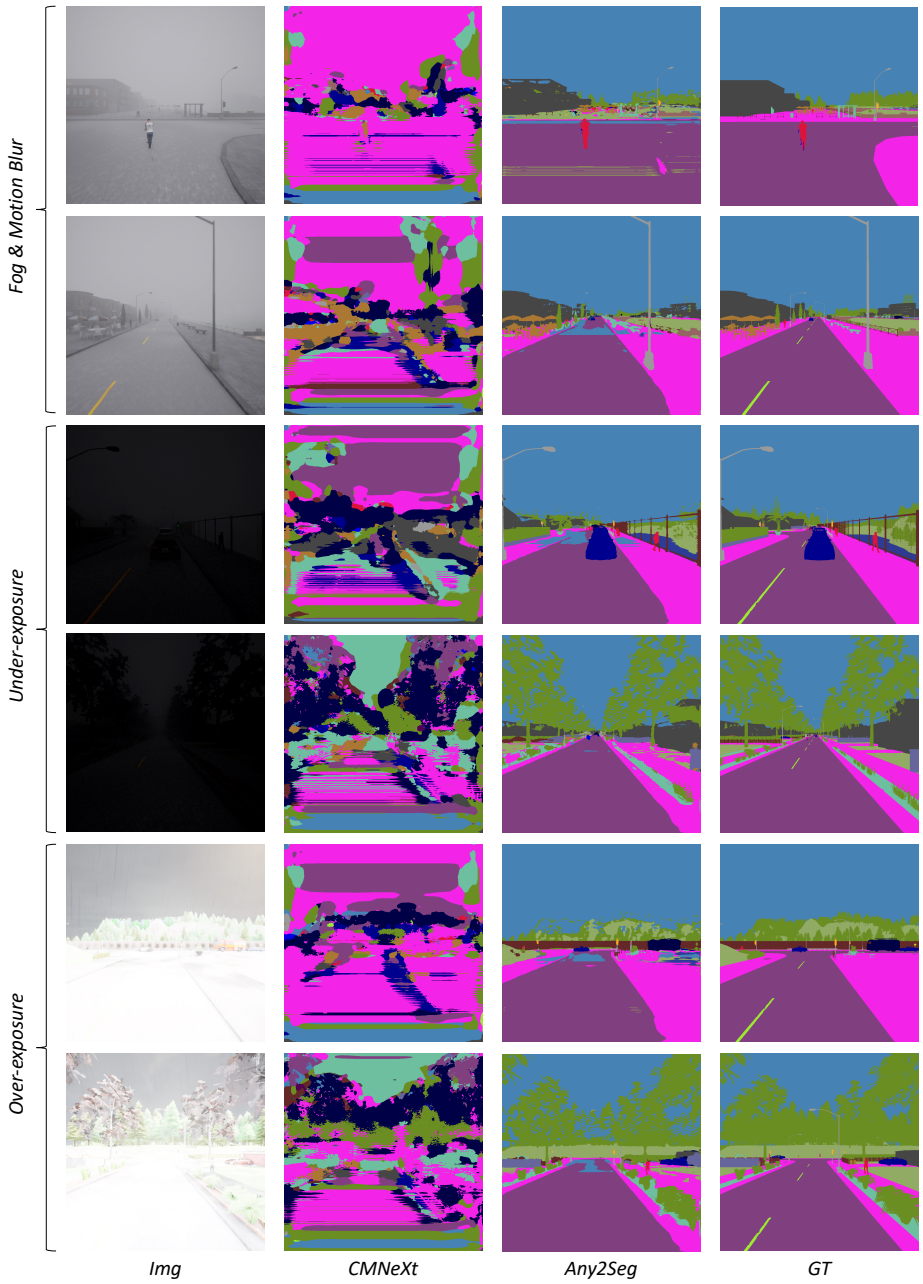

Fig. 1: System-level Modality-Incomplete Semantic Segmentation (MISS) validation results on DEVLIER Dataset with only depth data input.

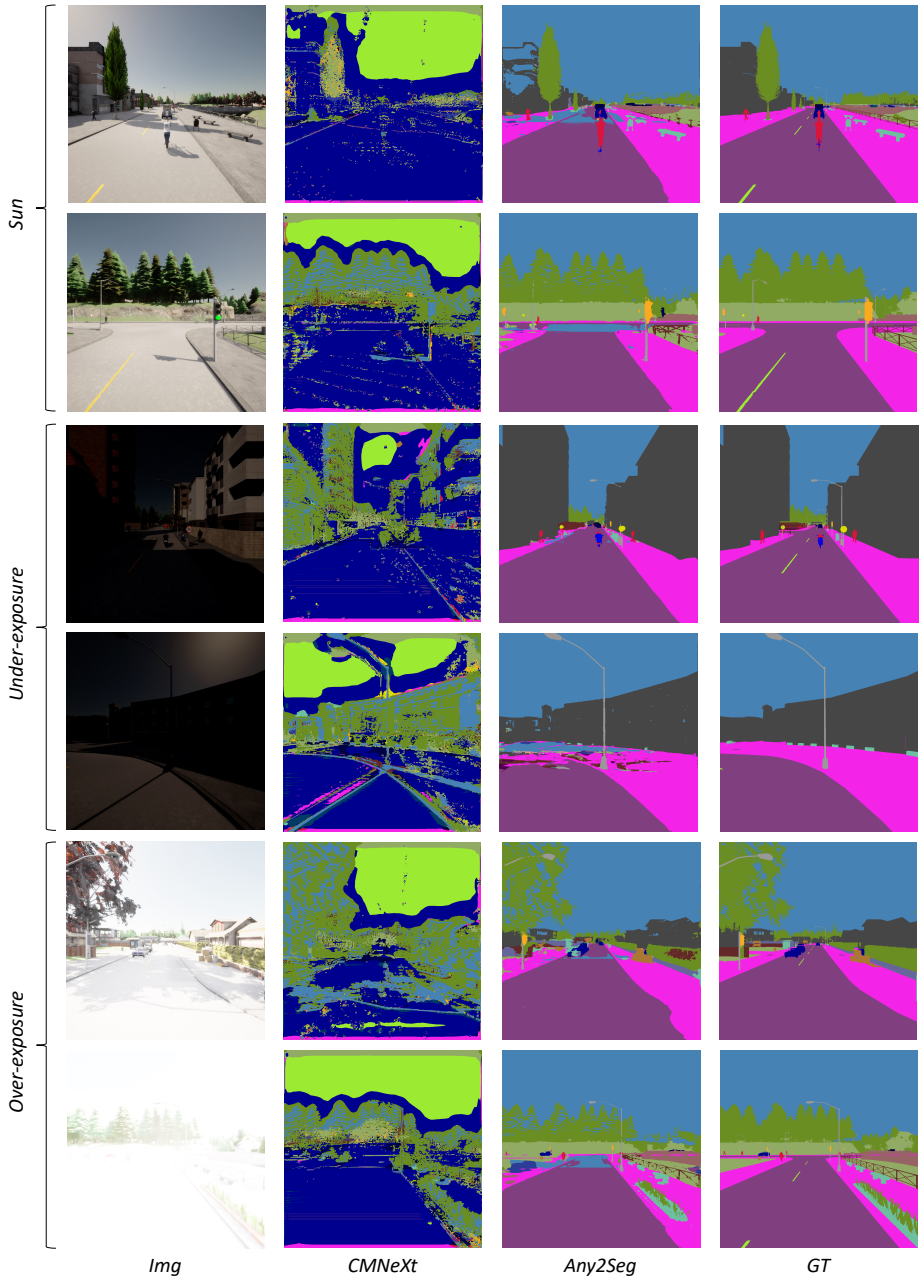

Fig. 2: System-level Modality-Incomplete Semantic Segmentation (MISS) validation results on DEVLIER Dataset with depth and Event data input.

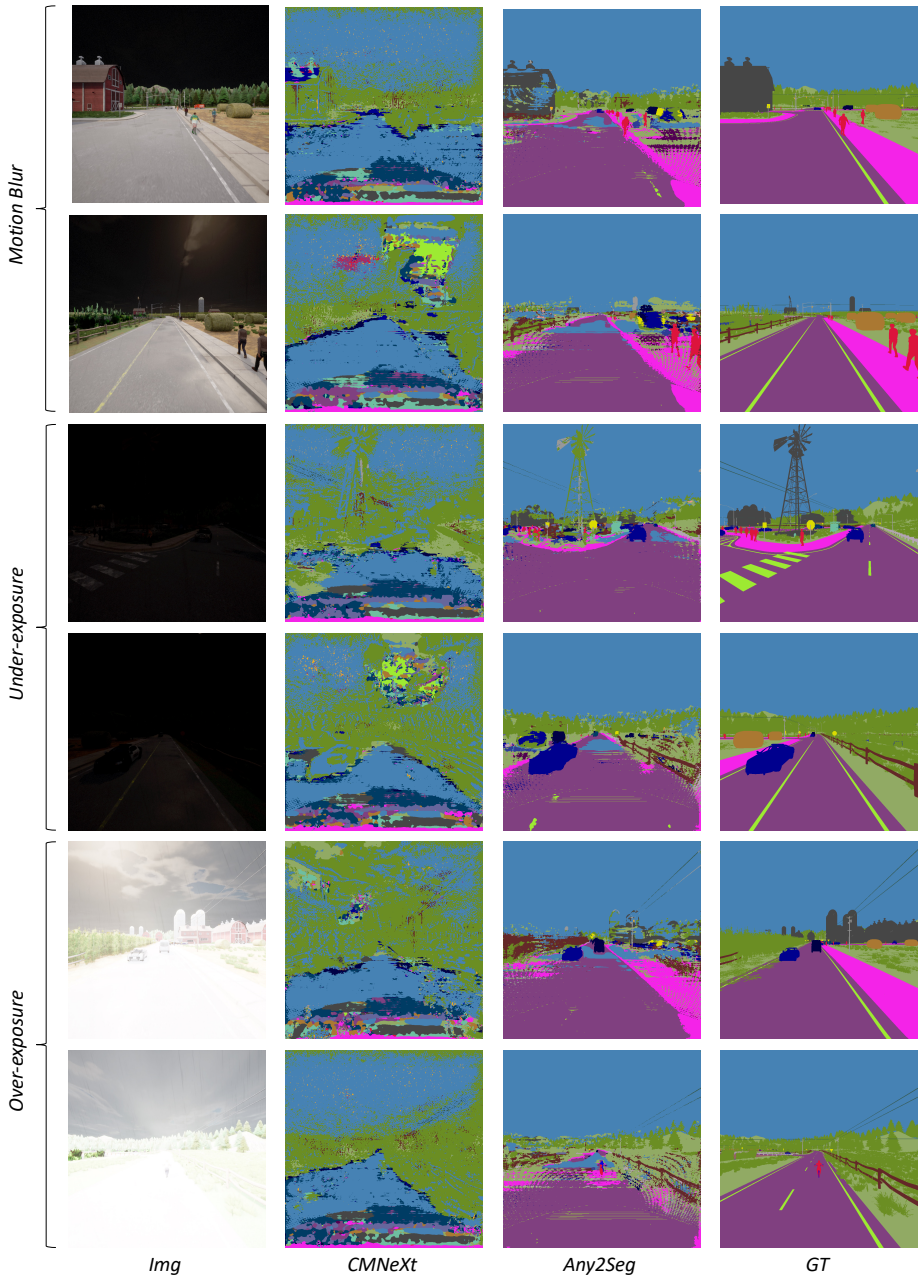

Fig. 3: System-level Modality-Incomplete Semantic Segmentation (MISS) validation results on DEVLIER Dataset with depth, event, and LiDAR data input.

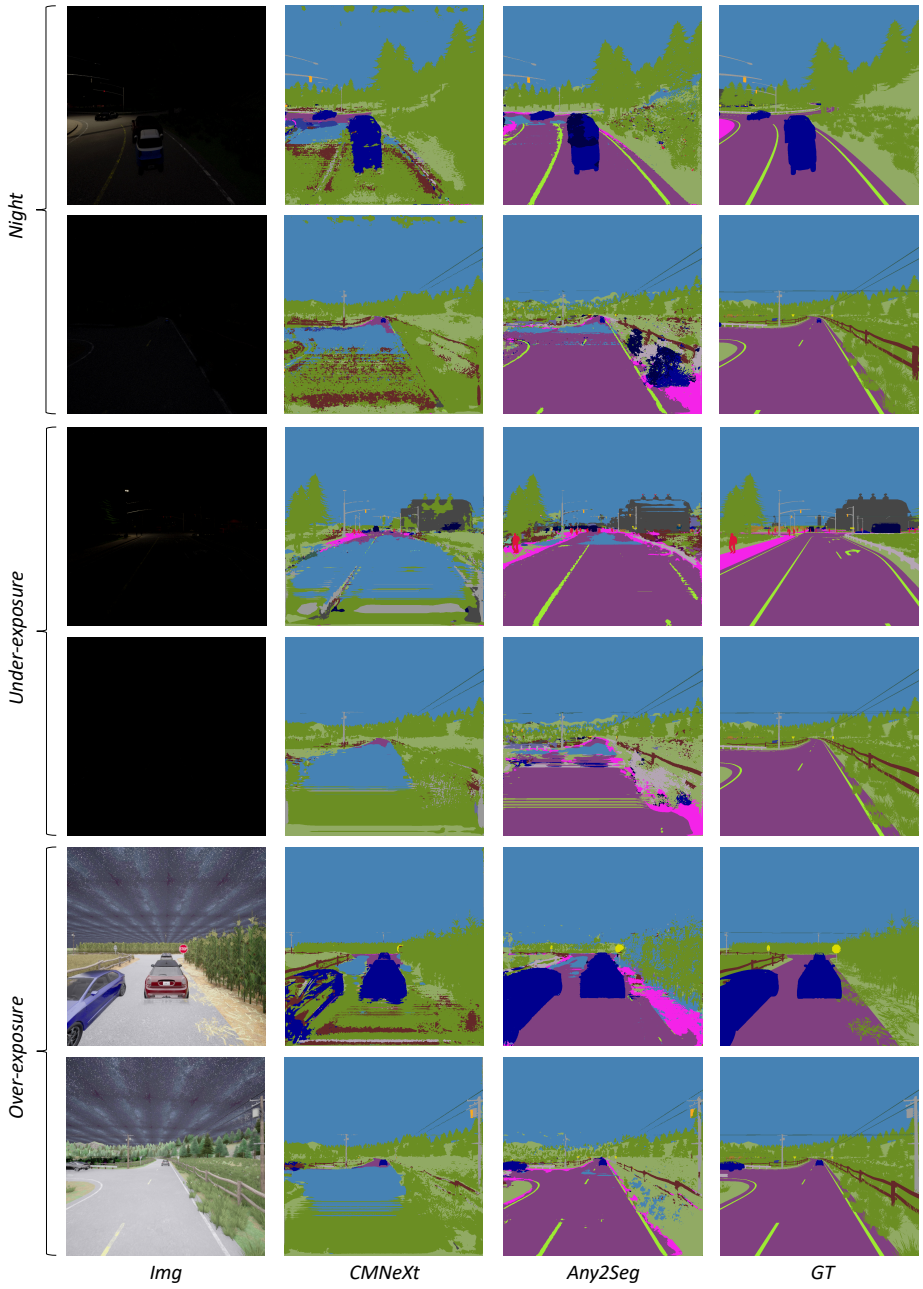

Fig. 4: System-level Modality-Incomplete Semantic Segmentation (MISS) validation results on DEVLIER Dataset with RGR and depth data input.

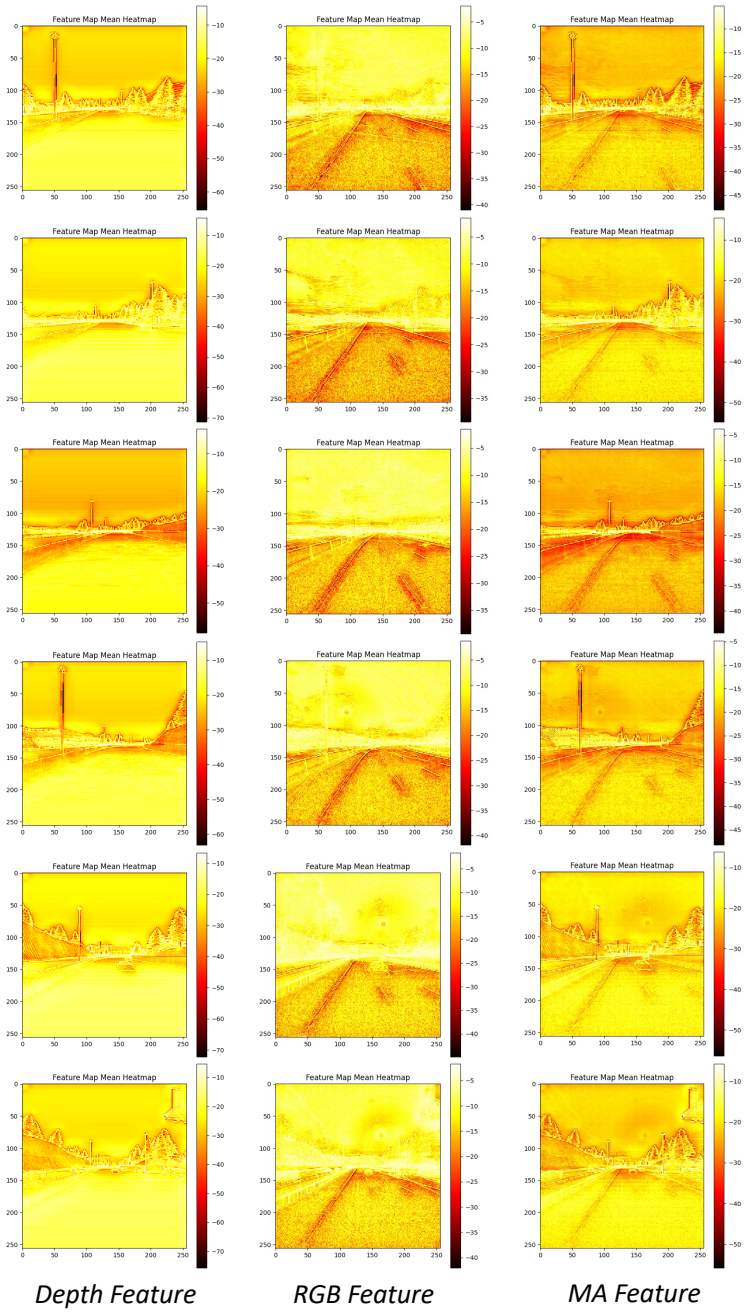

Fig. 5: Visualization of multi-modal features under different conditions on DEVLIER. MA feature: Modality-agnostic feature.

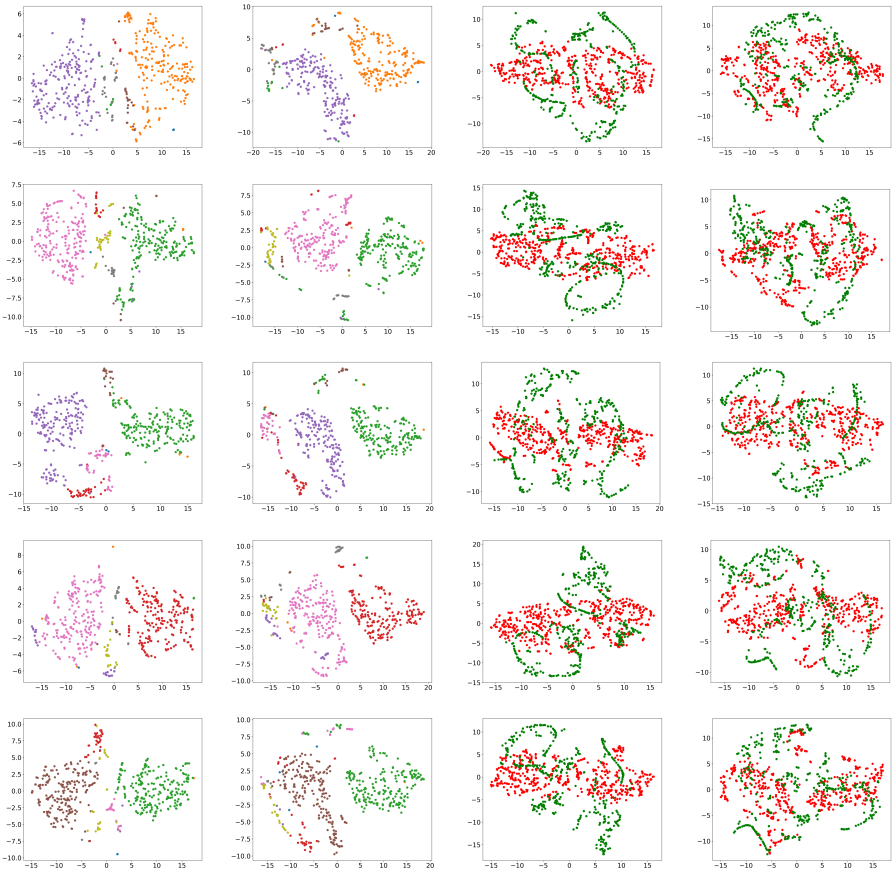

Fig. 6: t-SNE visualization. Red and green points in (c) and (d) represent RGB and Depth features, respectively. In (a) and (b), colors indicate distinct semantic classes.

References

1. 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)
2. 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)