Reducing Unimodal Bias in Multi-Modal Semantic Segmentation with Multi-Scale Functional Entropy Regularization –Supplementary Material–

Xu Zheng[1,2] Yuanhuiyi Lyu[1] Lutao Jiang[1]
Danda Pani Paudel[2] Luc Van Gool[2] Xuming Hu[1,3,*]

[1]AI Thrust, HKUST(GZ) [2]INSAIT, Sofia University "St. Kliment Ohridski" [3]CSE, HKUST

1. Related Work

1.1. Multi-modal Semantic Segmentation

Multi-modal Semantic Segmentation seeks to combine RGB with complementary modalities such as depth [4, 7, 8, 11, 12, 21–23, 27–30, 36, 54], thermal [6, 10, 18, 25, 26, 33, 34, 41, 42, 46, 56], events [1, 5, 38, 48, 55], and LiDAR [2, 13, 14, 19, 31, 35, 37, 57]. Advances in sensor technology have driven significant progress in multi-modal fusion [5, 16, 17, 21, 43, 44, 47, 49, 53], evolving from dual-modality to comprehensive multi-modal systems, like MCubeSNet [15], which improves scene understanding by utilizing richer sensor data.

From an architectural standpoint, multi-modal fusion models are generally classified into three categories: separate branches [3, 24, 32, 41], joint branches [7, 31], and asymmetric branches [39, 40]. A common approach in these models is to treat RGB as the primary modality, while auxiliary sensors provide additional information. For instance, CMNeXT [40] prioritizes RGB, incorporating other sensors to supplement the data. However, RGB alone may be insufficient, particularly in challenging conditions like low light or nighttime. This limitation highlights the need for more robust fusion models that leverage the strengths of multiple modalities, minimizing dependence on any single sensor. In this context, Liu *et al*. [20] introduced the concept of modality-incomplete scene segmentation, addressing deficiencies at both the system and sensor levels. In contrast to these works focused on architectural design, our approach in this paper aims to achieve balanced multi-modal training without introducing additional parameters. To accomplish this, we apply a plug-and-play regularization term to existing semantic segmentation backbones. The regularization term is achieved at both the high-level feature and output prediction levels, maximizing the potential of multi-scale segmentation backbones.

*Corresponding author.

1.2. Unimodal Bias

A key challenge in multi-modal tasks, especially for multi-modal semantic segmentation, is unimodal bias, where where models favor one modality over others, leading to suboptimal performance. Specifically, when the frame camera is included in the evaluation, the performance significantly increases. However, when it is missing, the performance drops sharply. This bias arises when models prioritize the most easily learnable or information-rich modality under specific conditions, neglecting the complementary benefits of other modalities. This can lead to a significant degradation in performance when the dominant modality is missing or unreliable [20, 50, 52], such as in cases of RGB corruption or thermal sensor noise, which can be particularly problematic for safety-critical applications like autonomous navigation. Several strategies have been proposed to address unimodal bias in multi-modal learning [50, 52]. Traditional multi-modal fusion models often do not explicitly regulate how each modality contributes to the final predictions, which can lead to over-reliance on a single input. Recent works, such as MAGIC [52] and Any2Seg [50] have sought to address this limitation by introducing well-designed training objectives.

In multi-modal learning, methods [45, 51] often draw from information theory, particularly the concept of functional entropy [9]. Entropy-based approaches quantify a model's uncertainty in its reliance on different modalities. A system with high entropy distributes its "attention" evenly across modalities, while low entropy signals an over-reliance on a single modality. Building on the application of functional Fisher information in visual question answering tasks [9], directly applying this approach to semantic segmentation presents challenges. Unlike visual question answering, which focuses on interpreting questions based on a single modality, segmentation tasks require the integration of both spatial and contextual information across multiple modalities. To address this, we introduce multi-scale reg-

ularization terms at both the feature and prediction levels. These regularization terms provide a principled framework to mitigate unimodal bias, encouraging a more balanced utilization of all available sensors. This ensures that models maintain accuracy even in scenarios where certain modalities degrade or fail—an essential consideration in environments with unreliable or variable sensor performance.

2. Experimental Details.

2.1. Datasets

DELIVER [40] is a large-scale multi-modal segmentation dataset which includes Depth, LiDAR, Views, Event, RGB data, based on the CARLA simulator. DELIVER [40] 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) of 20 category segmentation annotations. It has 302/96/102 image pairs for training/validation/testing at the size of 1224 $\times$ 1024.

2.2. Implementation Details.

We train our method on 8 $\times$ H100 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 [16, 40, 50, 51]. ImageNet-1K pre-trained weight is used as the pre-trained weight.

2.3. 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 dividing the intersection of the two segmentation maps by their union. IoU 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.

3. Additional Experimental Results

We evaluate the performance of our method on the real-world DELIVER dataset, focusing on dual-modality fusion in semantic segmentation tasks. Table 1 summarizes the validation results using two modality combinations: RGB-Depth and RGB-Event. For the RGB-Depth fusion task, our method achieves a mean Intersection over Union (IoU) of 60.37%, outperforming both CMNeXt (22.81%) and MAGIC (54.39%) by a significant margin. Notably, our approach yields the best performance when combining both modalities (66.46%) and shows a substantial improvement in RGB modality (55.04%) compared to MAGIC (37.26%). Although the improvement in Depth modality (59.60%) is smaller than MAGIC's (59.02%), the overall performance boost over the state-of-the-art (SoTA) methods is +5.98.

In the RGB-Event fusion task, our method achieves a mean IoU of 47.42%, which is higher than CMNeXt (21.92%) and MAGIC (43.76%). While our method shows a slight drop in RGB modality (56.90% vs. MAGIC's 58.00%), it outperforms MAGIC significantly in the Event modality (29.21% vs. 14.81%), resulting in a combined performance of 56.06, which is just -2.42% below MAGIC's 58.48%. The overall improvement over SoTA in this case is +3.66%. These results demonstrate the robustness and effectiveness of our approach in handling multi-modal data, providing more balanced and improved segmentation performance compared to current state-of-the-art methods.

We evaluate the performance of our method on the real-world MCubeS dataset for dual-modality semantic segmentation. Table 2 presents the validation results for three different modality combinations: Image-Aolp, Image-Dolp, and Image-Nolp. For the Image-Aolp fusion task, our method achieves a mean IoU of 46.24%, outperforming both CMNeXt (11.68%) and MAGIC (34.49%). Specifically, our method yields a combined performance of 50.65, with a significant improvement in the Aolp modality (39.15%), whereas MAGIC's Aolp performance is 0.27%. The overall improvement over state-of-the-art methods is +11.75%, with our method outperforming MAGIC in the Aolp modality by +38.88%. In the Image-Dolp fusion task, our method achieves a mean IoU of 39.89%, surpassing CMNeXt (12.00%) and MAGIC (33.32%). Our method shows strong performance in both Image (48.23%) and Dolp (21.13%) modalities, with a combined IoU of 50.31. This results in an improvement of +6.57 over the state-of-the-art. For the Image-Nolp fusion task, our

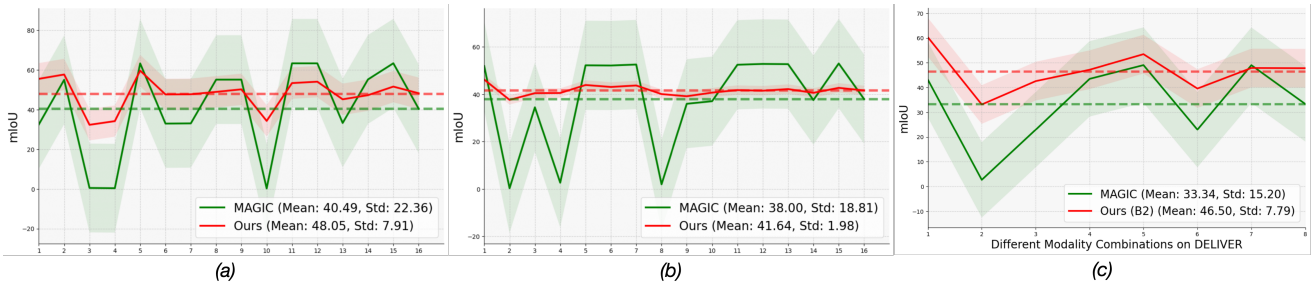

Figure 1. Balanced multi-modal performance comparison between ours and MAGIC [52] on (a) DELIVER, (b) MCubeS, and (c) MUSES.

Table 1. Results of anymodal semantic segmentation validation with dual modalities on real-world benchmark DELIVER dataset.

Method	Publication	RGB-Depth Fusion			Mean
		RGB	Dpeth	Both	
CMNeXt	CVPR 2023	1.60	1.44	63.58	22.81
MAGIC	ECCV 2024	37.26	59.02	66.89	54.39
Ours	-	55.04	59.60	66.46	60.37
w.r.t. SoTA	-	+17.58	+0.58	-0.43	**+5.98**
Method	Publication	RGB-Event Fusion			Mean
		RGB	Event	Both	
CMNeXt	CVPR 2023	4.82	3.45	57.48	21.92
MAGIC	ECCV 2024	58.00	14.81	58.48	43.76
Ours	-	56.90	29.21	56.06	47.42
w.r.t. SoTA	-	-1.10	+14.40	-2.42	**+3.66**

Table 2. Results of anymodal semantic segmentation validation with dual modalities on real-world benchmark MCubeS dataset.

Method	Publication	Image-Aolp Fusion			Mean
		Image	Aolp	Both	
CMNeXt	CVPR 2023	3.99	1.74	29.31	11.68
MAGIC	ECCV 2024	51.45	0.27	51.45	34.49
Ours	-	48.93	39.15	50.65	46.24
w.r.t. SoTA	-	-2.52	+38.88	-0.80	**+11.75**
Method	Publication	Image-Dolp Fusion			Mean
		Image	Dolp	Both	
CMNeXt	CVPR 2023	2.26	0.71	33.02	12.00
MAGIC	ECCV 2024	49.93	0.06	49.96	33.32
Ours	-	48.23	21.13	50.31	39.89
w.r.t. SoTA	-	-0.70	+21.07	+0.35	**+6.57**
Method	Publication	Image-Nolp Fusion			Mean
		Image	Nolp	Both	
CMNeXt	CVPR 2023	2.14	1.53	33.39	12.35
MAGIC	ECCV 2024	51.20	3.03	51.69	35.31
Ours	-	48.49	16.63	51.17	38.76
w.r.t. SoTA	-	-2.71	+13.60	-0.52	**+3.45**

method achieves a mean IoU of 38.76, again outperforming CMNeXt (12.35%) and MAGIC (35.31%). Our method shows solid performance in the Nolp modality (16.63%) and combines well with the Image modality (48.49%), resulting in a combined IoU of 51.17%, with an improvement of +3.45% over state-of-the-art methods. These results demonstrate the effectiveness of our approach, providing substantial improvements in dual-modality semantic segmentation tasks across multiple modality combinations. Our method consistently outperforms existing methods, offering a more balanced and robust segmentation performance on the MCubeS dataset.

4. Additional Ablation Study

Unimodal Bias As shown in Fig.1, our method demonstrates improved performance compared to other methods, such as MAGIC[52], which aim to achieve balanced cross-modal performance. Specifically, while MAGIC (represented by the green line) exhibits notable unimodal bias, with large fluctuations in performance across different modality combinations, our method (represented by the red line) shows more consistent and stable performance. MAGIC has an average mIoU of 40.49% with a high standard deviation of 22.36%, indicating significant sensitivity to certain modality pairs. In contrast, our method achieves a higher mean mIoU of 48.05% with a much smaller standard deviation of 7.91%, highlighting its robustness across varying modality combinations. This trend further underscores the absence of unimodal bias in our approach, as it handles all modality combinations effectively. On the other hand, MAGIC suffers from considerable fluctuations, particularly in some modality pairs, as seen in the figure. The shaded areas around the curves further emphasize the reduced variability in our method, demonstrating its stability and reliability in achieving balanced performance across different

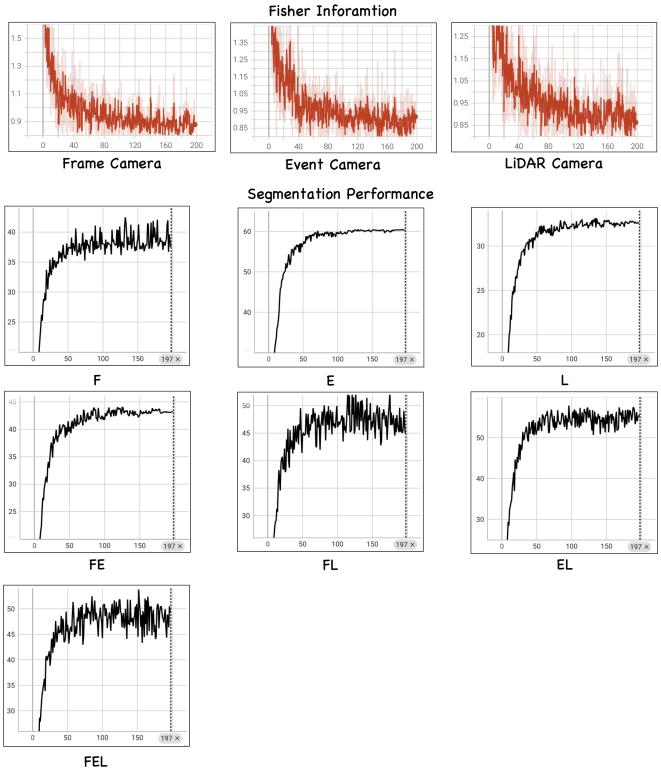

Figure 2. Functional fisher information and segmentation performance across training with our proposed regularization terms on MUSES dataset.

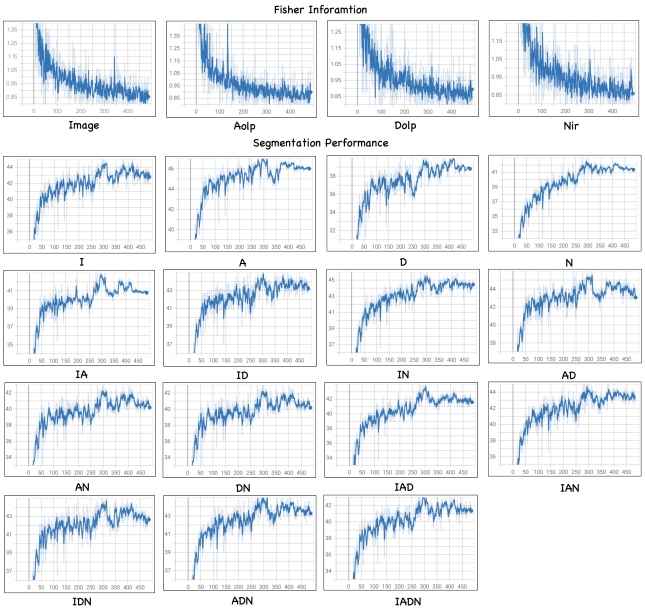

Figure 3. Functional fisher information and segmentation performance across training with our proposed regularization terms on MCubeS dataset.

modalities.

Fisher Information and Performance across Training Fig. 2 shows the segmentation performance across training with our proposed regularization terms on the DELIVER dataset. The plots demonstrate the impact of the multi-scale regularization terms on model performance with different modality combinations, measured by the evaluation score (mIoU) at each training step. This ablation study demonstrates that the proposed regularization terms leads to consistent improvements in segmentation performance across any modality combinations. Particularly, the use of multi-scale regularization terms results in more stable and progressively better performance across training, reinforcing the effectiveness of our approach. Meanwhile, Fig. 2 and Fig. 3 show similar results on the MUSES and MCubeS datasets, respectively. These findings across multiple datasets and modality combinations confirm that our proposed regularization terms, particularly when applied at multiple scales, significantly enhance model stability and overall segmentation performance.

Qualitative Results Figure 4 presents qualitative results comparing the performance of our method with the baseline methods CMNeXt and MAGIC under various weather conditions: cloudy, night, rain, and sun. For each condition, the results show the input image, the event data, the ground truth (GT), and the predictions using different regularization terms: RDEL, DEL, EL, and E.

- **Cloudy Weather**: In this condition, CMNeXt and MAGIC struggle with some occlusions and areas with poor visibility, as seen in the noisy predictions (e.g., in the middle and lower parts of the road). Our method, on the other hand, shows significantly cleaner and more accurate segmentation, particularly when using the RDEL and DEL combinations, providing consistent road and vehicle segmentation.

- **Night Condition**: At night, CMNeXt fails to properly segment the road and vehicles, with much of the road appearing unclear in the predictions. MAGIC also shows similar challenges, especially in low-light areas. Our method, particularly with DEL and RDEL regularization, performs better by handling the lighting variation more effectively and maintaining the structure of the road and vehicle features.

- **Rain Weather**: Under rainy conditions, CMNeXt and MAGIC both struggle with poor visibility, resulting in over-segmented or under-segmented areas in the road and surroundings. Our method shows a clear improvement in handling the noise introduced by rain, providing more accurate segmentations, especially when using the DEL and EL regularization terms.

- **Sun Condition**: In bright sunlight, CMNeXt and MAGIC still show poor segmentation quality, with road features and vehicles becoming poorly defined. Our method, however, successfully segments the road and vehicles un-

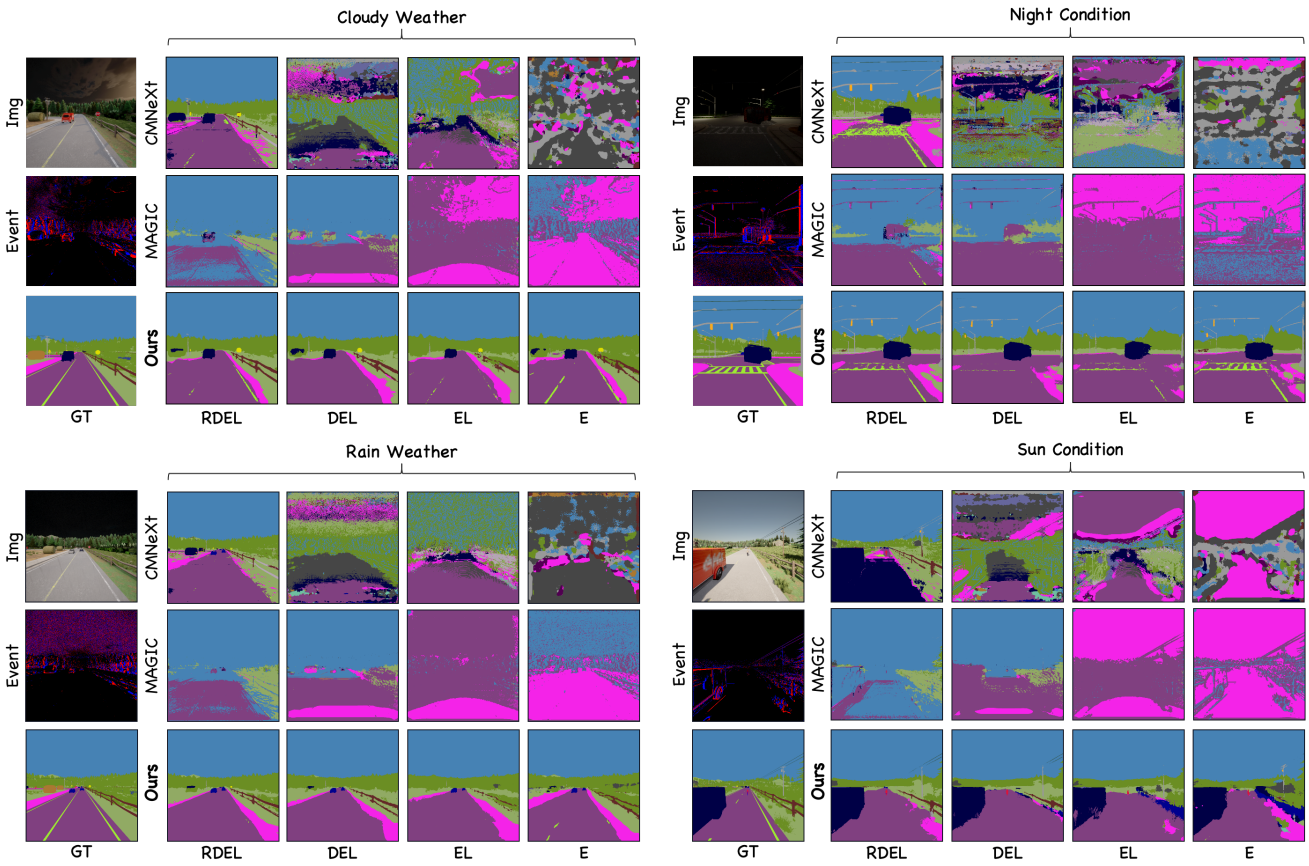

Figure 4. Comparison.

der challenging lighting conditions, with RDEL and DEL providing the most robust segmentation results.

Overall, our method outperforms both CMNeXt and MAGIC in all weather conditions, providing more accurate and stable segmentation. The use of multi-modal data and regularization terms, particularly RDEL and DEL, plays a crucial role in achieving consistent segmentation performance across varying environmental factors.

References

[1] Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *Proceedings of IEEE/CVF CVPR Workshops*, pages 0–0, 2019. 1

[2] Shubhankar Borse, Marvin Klingner, Varun Ravi Kumar, Hong Cai, Abdulaziz Almuzairee, Senthil Yogamani, and Fatih Porikli. X-align: Cross-modal cross-view alignment for bird's-eye-view segmentation. In *Proceedings of the IEEE/CVF WACV*, pages 3287–3297, 2023. 1

[3] Tim Broedermann, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Hrfuser: A multi-resolution sensor fusion architecture for 2d object detection. In *IEEE International Conference on Intelligent Transportation Systems*, pages 4159–4166, 2023. 1

[4] Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF ICCV*, pages 7088–7097, 2021. 1

[5] Jiahang Cao, Xu Zheng, Yuanhuiyi Lyu, Jiaxu Wang, Renjing Xu, and Lin Wang. Chasing day and night: Towards robust and efficient all-day object detection guided by an event camera. *arXiv:2309.09297*, 2023. 1

[6] Gang Chen, Feng Shao, Xiongli Chai, Hangwei Chen, Qiuping Jiang, Xiangchao Meng, and Yo-Sung Ho. Modality-induced transfer-fusion network for rgb-d and rgb-t salient object detection. *IEEE TCSVT*, 33(4):1787–1801, 2022. 1

[7] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng. Spatial information guided convolution for real-time rgbd semantic segmentation. *IEEE TIP*, 30:2313–2324, 2021. 1

[8] Runmin Cong, Qinwei Lin, Chen Zhang, Chongyi Li, Xiaochun Cao, Qingming Huang, and Yao Zhao. Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection. *IEEE TIP*, 31:6800–6815, 2022. 1

[9] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33:3197–3208, 2020. 1

[10] Tianrui Hui, Zizheng Xun, Fengguang Peng, Junshi Huang, Xiaoming Wei, Xiaolin Wei, Jiao Dai, Jizhong Han, and Si Liu. Bridging search region interaction with template for rgb-t tracking. In *Proceedings of IEEE/CVF CVPR*, pages 13630–13639, 2023. 1

[11] Wei Ji, Ge Yan, Jingjing Li, Yongri Piao, Shunyu Yao, Miao Zhang, Li Cheng, and Huchuan Lu. Dmra: Depth-induced multi-scale recurrent attention network for rgb-d saliency detection. *IEEE TIP*, 31:2321–2336, 2022. 1

[12] Minhyeok Lee, Chaewon Park, Suhwan Cho, and Sangyoun Lee. Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In *ECCV*, pages 630–647. Springer, 2022. 1

[13] Jiale Li, Hang Dai, Hao Han, and Yong Ding. Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21694–21704, 2023. 1

[14] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of IEEE/CVF CVPR*, pages 17182–17191, 2022. 1

[15] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of IEEE/CVF CVPR*, pages 19800–19808, 2022. 1

[16] Chenfei Liao, Kaiyu Lei, Xu Zheng, Junha Moon, Zhixiong Wang, Yixuan Wang, Danda Pani Paudel, Luc Van Gool, and Xuming Hu. Benchmarking multi-modal semantic segmentation under sensor failures: Missing and noisy modality robustness. *CoRR*, abs/2503.18445, 2025. 1, 2

[17] Chenfei Liao, Xu Zheng, Yuanhuiyi Lyu, Haiwei Xue, Yihong Cao, Jiawen Wang, Kailun Yang, and Xuming Hu. Memorysam: Memorize modalities and semantics with segment anything model 2 for multi-modal semantic segmentation. *CoRR*, abs/2503.06700, 2025. 1

[18] Guibiao Liao, Wei Gao, Ge Li, Junle Wang, and Sam Kwong. Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE TCSVT*, 32 (11):7646–7661, 2022. 1

[19] Haisong Liu, Tao Lu, Yihui Xu, Jia Liu, Wenjie Li, and Lijun Chen. Camliflow: Bidirectional camera-lidar fusion for joint optical flow and scene flow estimation. In *Proceedings of IEEE/CVF CVPR*, pages 5791–5801, 2022. 1

[20] Ruiping Liu, Jiaming Zhang, Kunyu Peng, Yufan Chen, Ke Cao, Junwei Zheng, M Saquib Sarfraz, Kailun Yang, and Rainer Stiefelhagen. Fourier prompt tuning for modality-incomplete scene segmentation. In *IEEE Intelligent Vehicles Symposium*, pages 961–968, 2024. 1

[21] Yuanhuiyi Lyu, Xu Zheng, Dahun Kim, and Lin Wang. Omnibind: Teach to build unequal-scale modality interaction for omni-bind of all. *arXiv:2405.16108*, 2024. 1

[22] Yuanhuiyi Lyu, Xu Zheng, and Lin Wang. Image anything: Towards reasoning-coherent and training-free multi-modal image generation. *arXiv:2401.17664*, 2024.

[23] Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. Unibind: Llm-augmented unified and balanced representation space to bind them all. In *Proceedings of IEEE/CVF CVPR*, pages 26752–26762, 2024. 1

[24] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Bev-guided multi-modality fusion for driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21960–21969, 2023. 1

[25] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Caver: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE TIP*, 32:892–904, 2023. 1

[26] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE ICRA*, pages 9441–9447. IEEE, 2020. 1

[27] Mengke Song, Wenfeng Song, Guowei Yang, and Chenglizhao Chen. Improving rgb-d salient object detection via modality-aware decoder. *IEEE TIP*, 31:6124–6138, 2022. 1

[28] Fengyun Wang, Jinshan Pan, Shoukun Xu, and Jinhui Tang. Learning discriminative cross-modality features for rgb-d saliency detection. *IEEE TIP*, 31:1285–1297, 2022.

[29] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *NeurIPS*, 33:4835–4845, 2020.

[30] Yikai Wang, Fuchun Sun, Ming Lu, and Anbang Yao. Learning deep multimodal feature representation with asymmetric multi-layer fusion. In *Proceedings of the 28th ACM MM*, pages 3902–3910, 2020. 1

[31] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12186–12195, 2022. 1

[32] Shicai Wei, Chunbo Luo, and Yang Luo. 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*, pages 20039–20049, 2023. 1

[33] Wei Wu, Tao Chu, and Qiong Liu. Complementarity-aware cross-modal feature fusion network for rgb-t semantic segmentation. *Pattern Recognition*, 131:108881, 2022. 1

[34] Zhengxuan Xie, Feng Shao, Gang Chen, Hangwei Chen, Qiuping Jiang, Xiangchao Meng, and Yo-Sung Ho. Cross-modality double bidirectional interaction and fusion network for rgb-t salient object detection. *IEEE TCSVT*, 2023. 1

[35] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *ECCV*, pages 677–695. Springer, 2022. 1

[36] Xiaowen Ying and Mooi Choo Chuah. Uctnet: Uncertainty-aware cross-modal transformer network for indoor rgb-d semantic segmentation. In *ECCV*, pages 20–37. Springer, 2022. 1

[37] Boxiang Zhang, Zunran Wang, Yonggen Ling, Yuanyuan Guan, Shenghao Zhang, and Wenhui Li. Mx2m: Masked cross-modality modeling in domain adaptation for 3d semantic segmentation. In *Proceedings of the AAAI*, pages 3401–3409, 2023. 1

[38] Jiaming Zhang, Kailun Yang, and Rainer Stiefelhagen. Issafe: Improving semantic segmentation in accidents by fusing event-based data. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1132–1139. IEEE, 2021. 1

[39] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *arXiv:2203.04838*, 2022. 1

[40] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023. 1, 2

[41] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. 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*, pages 2633–2642, 2021. 1

[42] Tianlu Zhang, Hongyuan Guo, Qiang Jiao, Qiang Zhang, and Jungong Han. Efficient rgb-t tracking via cross-modality distillation. In *Proceedings of IEEE/CVF CVPR*, pages 5404–5413, 2023. 1

[43] Weiming Zhang, Yexin Liu, Xu Zheng, and Lin Wang. Goodsam: Bridging domain and capacity gaps via segment anything model for distortion-aware panoramic semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 28264–28273. IEEE, 2024. 1

[44] Weiming Zhang, Yexin Liu, Xu Zheng, and Lin Wang. Goodsam++: Bridging domain and capacity gaps via segment anything model for panoramic semantic segmentation. *CoRR*, abs/2408.09115, 2024. 1

[45] Yedi Zhang, Peter E Latham, and Andrew M Saxe. Understanding unimodal bias in multimodal deep linear networks. In *International Conference on Machine Learning*, 2024. 1

[46] Jiayi Zhao, Fei Teng, Kai Luo, Guoqiang Zhao, Zhiyong Li, Xu Zheng, and Kailun Yang. Unveiling the potential of segment anything model 2 for rgb-thermal semantic segmentation with language guidance. *CoRR*, abs/2503.02581, 2025. 1

[47] Xu Zheng and Lin Wang. Eventdance++: Language-guided unsupervised source-free cross-modal adaptation for event-based object recognition. *CoRR*, abs/2409.12778, 2024. 1

[48] Xu Zheng and Lin Wang. 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*, pages 17448–17458, 2024. 1

[49] Xu Zheng, Yuanhuiyi Lyu, Lutao Jiang, Jiazhou Zhou, Lin Wang, and Xuming Hu. MAGIC++: efficient and resilient modality-agnostic semantic segmentation via hierarchical modality selection. *CoRR*, abs/2412.16876, 2024. 1

[50] Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Learning modality-agnostic representation for semantic segmentation from any modalities. In *European Conference on Computer Vision*, pages 146–165. Springer, 2024. 1, 2

[51] Xu Zheng, Haiwei Xue, Jialei Chen, Yibo Yan, Lutao Jiang, Yuanhuiyi Lyu, Kailun Yang, Linfeng Zhang, and Xuming Hu. Learning robust anymodal segmentor with unimodal and cross-modal distillation. *arXiv preprint arXiv:2411.17141*, 2024. 1, 2

[52] Xu Zheng, Yuanhuiyi Lyu, Jiazhou Zhou, and Lin Wang. Centering the value of every modality: Towards efficient and resilient modality-agnostic semantic segmentation. In *ECCV*, pages 192–212. Springer, 2025. 1, 3

[53] Ding Zhong, Xu Zheng, Chenfei Liao, Yuanhuiyi Lyu, Jialei Chen, Shengyang Wu, Linfeng Zhang, and Xuming Hu. Omnisam: Omnidirectional segment anything model for UDA in panoramic semantic segmentation. *CoRR*, abs/2503.07098, 2025. 1

[54] Hao Zhou, Lu Qi, Zhaoliang Wan, Hai Huang, and Xu Yang. Rgb-d co-attention network for semantic segmentation. In *Proceedings of the ACCV*, 2020. 1

[55] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. 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*, pages 18633–18643, 2024. 1

[56] Wujie Zhou, Han Zhang, Weiqing Yan, and Weisi Lin. Mmsmcnet: Modal memory sharing and morphological complementary networks for rgb-t urban scene semantic segmentation. *IEEE TCSVT*, 2023. 1

[57] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *Proceedings of the IEEE/CVF ICCV*, pages 16280–16290, 2021. 1