Decoupled and Reusable Adaptation for Efficient Cross-Modal Transfer

Supplementary Material

This supplementary material includes the following content:

- Sources of unlabeled dataset;
- More detailed training settings;
- More results and analyses;

1. More Experimental Details

1.1. Unlabeled Dataset

In Stage 1, we collect unlabeled datasets that covers diverse scenes, tasks, viewpoints, and illumination conditions for each target modality, and each is constructed from multiple public sources. The details are illustrated in Table 2. Since many datasets consist of frames extracted from videos (such as object tracking datasets), we sample only a subset of them according to the frame rate to ensure the diversity of training scenes. And the total numbers of unlabeled datasets for thermal, depth, polarization and event are at 100K, 150K, 50K and 100K respectively.

1.2. Detailed Training Setting

In Stage 1 training, we employ the original decoder in MAE as the MAE head and employ the random mask strategy with a mask ratio of 0.75. The student/teacher heads in DINOv2 training are three-layer MLPs and the output dim is set to 256. The teacher momentum of EMA in DINOv2 is set to 0.994 and is updated at the end of every training step. During MAE-guided consistency learning, the shared mask ratio of MAE and DINOv2 is randomly set to [0.1, 0.6]. And the filter ratio of the unlabeled dataset is set to 0.3 in terms of reconstruction error. The λ in collaborative training is set to 0.5. The data augmentation techniques include random horizontal flipping, random Gaussian blurring, random resized cropping and random color jittering. In Stage 2 training, the generalizable and modality specific experts are all standard Feed Forward Networks (FFNs). We employ a generalizable modality expert and four modality-specific experts in all of our experiments. The gating network is a two-layer MLP. The task head for semantic segmentation in our experiments is the same as CMNeXt [44]. During training in Stage 1 and Stage 2, the image size is standardized to 1024×1024 to match the input size of Hiera. The batch size for all downstream tasks is set to 4. For SUN-RGBD [31], the learning rate is set to 2e-4 with 250 training epochs. For NYU Depth V2 [29], the learning rate is 3.5e-4 with 300 training epochs. For MFNet [7], the learning rate is 6e-5 with 200 training epochs. For PST900 [28], the learning rate is also 6e-5 with 200 training epochs. For MCubeS [16], the learning rate is 4e-4 with 300 training epochs. For DELIVER [45], the learning rate is 6e-4 with 200 training epochs. All is scheduled by the Warmup Polynomial strategy with power 0.9. The first 10 epochs are to warm-up models with $0.1\times$ the original learning rate.

Table 1. More experiments and comparison on three different tasks with DINOv2.

RGB-T salient object detection task on VT5000					
Method	Backbone	S_m	E_m	F_β	MAE ↓
$\text{TPSSCL}_{\text{AAAI26}}$ [8]	Swin-B	0.922	0.958	0.902	-
$\text{SAMSOD}_{\text{TMM26}}$[23]	Hiera-T	0.923	0.964	0.921	0.021
Ours	ViT-B/14	**0.928**	**0.966**	**0.925**	**0.020**
RGB-T visual place recognition task on STheReO with R@1					
Method	Backbone	SUN-d	SUN-n	Vally-d	Vally-n
$\text{MMVPR}_{\text{IROS25}}$[42]	ViT-B/14	96.8	94.3	99.2	96.2
Ours	ViT-B/14	**97.9**	**95.8**	**99.4**	**97.4**
RGB-D 3D object detection task on Omni3D-SUN RGB-D					
Method	Backbone	AP25	AR25	AP50	AR50
$\text{CuTR}_{\text{CVPR25}}$ [11]	ViT-B/16	30.3	60.2	13.6	29.0
Ours	ViT-B/14	**31.2**	**61.7**	**15.3**	**31.1**

2. More Results and Analyses

2.1. Validation on DINOv2 and three more tasks

To strengthen the claims of our method that adapting to a broad sense foundation models and being task-agnostic to downstream scenarios, we further provide DINOv2 results on thermal and depth modalities, with evaluation on three different downstream tasks in Table 1.

Similarly to SAM2-based training pipeline, in Stage 1, we obtain the modality LoRAs for thermal and depth modalities on DINOv2 of ViT-B/14 encoder with the aforementioned unlabeled datasets. With the trained modality LoRAs, we conduct Stage 2 tuning. For RGB-T salient object detection task, we conduct experiments on VT5000 dataset and employ commonly used metrics including S-measure (S_m), F-measure (F_β), E-measure (E_m), and mean absolute error (MAE). We employ a U-Net style decoder and utilize the intermediate output features of blocks [3, 6, 9, 12] to construct feature pyramid. For RGB-T visual place recognition task, we employ the same dataset, metrics as well as the decoder structure as MMVPR [42]. For RGB-D 3D object detection task, we employ CuTR [11] as baseline, and the decoder structure and the evaluation metrics are also the same. Our experimental results demonstrate the adaptability of our transfer method to general-purpose VFMs and a diverse range of downstream tasks.

Table 2. Sources and the number of unlabeled data of each modality in Stage 1 training. In the dataset composed of video frames (such as object tracking tasks), we sample only a subset of frames according to the frame rate to ensure the diversity of training scenes.

Modality	**Thermal**	**Depth**	**Polarization**	**Event**
Total	100K	150K	50K	100K
Sources	FLIR [6] M3FD [20] LLVIP [10] KAIST [9] Freiburg [34] LSOTB-TIR [22] BU-TIV [38] RGBT234 [13] PTB-TIR [21] DVTOD [30] RGBT-Tiny [43]	KITTI-360 [18] ScanNet [4] Cityscapes [3] MegaDepth [14] DIODE [33] Synthia [26] Make3D [27] Matterport3D[1]	PolarRGB [41] PolarRR [12] RGBP-Glass [24] ZJU-RGBP [39] LLCP [40] PLIE [47] RGBP-Car [5] PGV-117 [25]	VisEvent [35] EventVOT [36] COESOT [32] MMPD [46] SDE [15] RLED [19] EvDET200K [37] EV-UAV [2]

Table 3. Cosine similarity between $\nabla\mathcal{L}_{\text{MAE}}$ and $\nabla\mathcal{L}_{\text{DINO}}$ during training.

	Depth			**Thermal**			**Polarization**			**Event**			
Epoch idx	0	1	2	0	1	2	0	1	2	0	1	2	Avg.
Cos. Sim.	0.25	0.30	0.27	0.18	0.20	0.23	0.13	0.12	0.15	0.23	0.20	0.20	**0.21**

2.2. Analysis on the loss synergy effect in PSST

Our model minimizes both reconstruction loss and semantic alignment loss and claim that non-RGB texture encodes semantics. To validate the synergy effect between the two objectives in non-RGB modalities, we compute the cosine sim between $\nabla\mathcal{L}_{\text{MAE}}$ and $\nabla\mathcal{L}_{\text{DINO}}$ to measure optimization conflict in Table 3. The average is 0.21, indicating that the two gradients are positively correlated rather than conflicting (which manifests as negative values). This synergy is also validated by our ablation study in Table 5, with a 3.3% drop when optimizing the two separately compared to joint training.

2.3. Analysis on the rank of modality LoRA

The rank of LoRA affects both the model's capacity to represent modality-specific knowledge and its parameter efficiency. Therefore, we compare different ranks to analyze their impact on the number of trainable parameters and the model's downstream performance. As shown in Table 4, we conduct experiments on the Thermal modality and evaluate performance variations on the PST900 [28] task. When $r = 8$, although the number of trainable parameters is only 0.3M, the LoRA fails to effectively represent the target modality. Increasing the rank to $r = 16$ doubles the parameter count and leads to a clear performance improvement. At $r = 32$, the trainable parameters reach 1.3M, and the downstream performance increases significantly to 88.7%. Further increasing the rank to $r = 64$ results in 2.6M trainable parameters with only marginal gains of 0.2%. Therefore, considering the trade-off between model performance and parameter efficiency, we set $r = 32$ as the default rank.

Table 4. Analysis on the LoRA rank.

rank	**# Param. (M)**	**mIoU (%)**
8	0.3	83.7
16	0.6	86.3
32	1.3	88.7
64	2.6	88.9

Table 5. Effectiveness analysis on our modality LoRA. To ensure a fair comparison, the reported MemorySAM [17] mIoU performance (%) on DELIVER [45] and MCubeS [16] datasets is based on the reproduced results.

Method	**DELIVER**	**MCubeS**
MemorySAM	62.5	51.6
MemorySAM + our LoRA	64.7	53.5

2.4. Analysis on the modality LoRA

Effectiveness. To further validate the capability of our Stage 1 trained modality LoRAs in representing modality-general and task-agnostic knowledge, we integrate the trained Depth-LoRA and Polarization-LoRA into the MemorySAM [17] encoder, which shares the same encoder as

Table 6. Stage 1 training cost with 1024*1024 input size using GeForce RTX 3090Ti.

Modal	Epochs	GPU-h	FLOPs	Data
Thermal	12	38.8	3.40E	100K
Polar.	12	19.4	1.69E	50K
Depth	9	43.7	3.82E	150K
Event	15	48.6	4.25E	100K

ours, and all other training settings are kept identical to those of the original MemorySAM [17]. The results are presented in Table 5. MemorySAM [17] achieves performance improvements of 2.2% and 1.9% mIoU on the DELIVER [45] and MCubeS [16] datasets, respectively, demonstrating the generalizability of our approach. Although the improvements achieved using MemorySAM [17] are smaller than the gains of 3.8% and 4.9% delivered by our method on the two datasets, this is expected because our TP-MoME adopts a more lightweight design, containing only 8.3M trainable parameters compared with MemorySAM's 15M [17]. Consequently, the absence of Stage 1 has a greater impact on our approach. Nevertheless, these results still demonstrate that even a more complete task-oriented transfer pipeline can benefit further when enhanced with our LoRA.

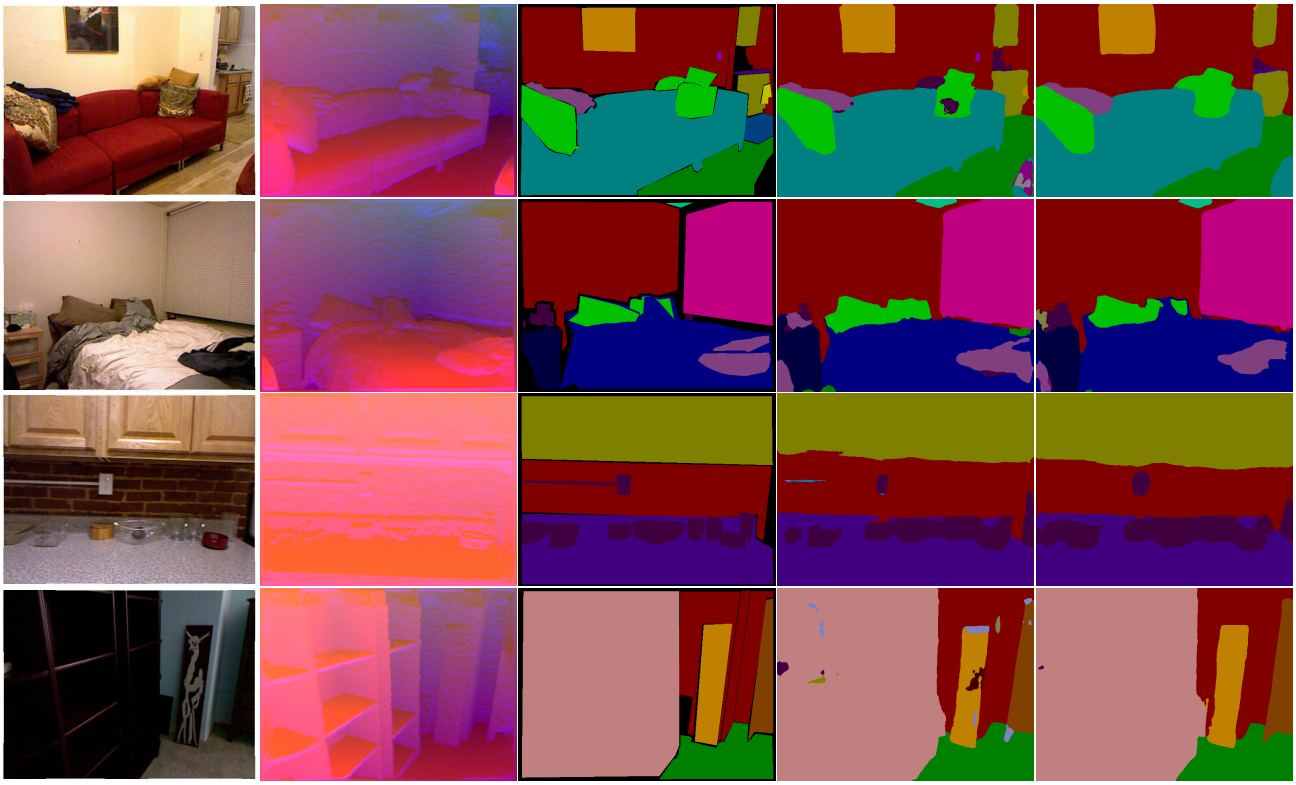
Generalization. With large-scale, heterogeneous target-modality data of diverse sensor types and environments, Stage 1 training enables modality LoRAs to capture modality-level invariant structural priors. The low-rank constraint also acts as an information bottleneck that filters out domain-specific variations, promoting generalizable features. This is empirically validated by: a single Depth LoRA performs consistently well on NYU (Kinect/indoor/real), SUN (Xtion/indoor/real) and DELIVER (outdoor/synthetic); a single Thermal LoRA generalizes across PST900 (FLIR/underground mine) and MFNet (InfRec R500/urban street) without negative transfer.

2.5. Stage 1 training cost

We quantify one-time training cost of Stage 1 in Table 6, including total training epochs, training time (GPU-hour), total FLOPs and data scale.

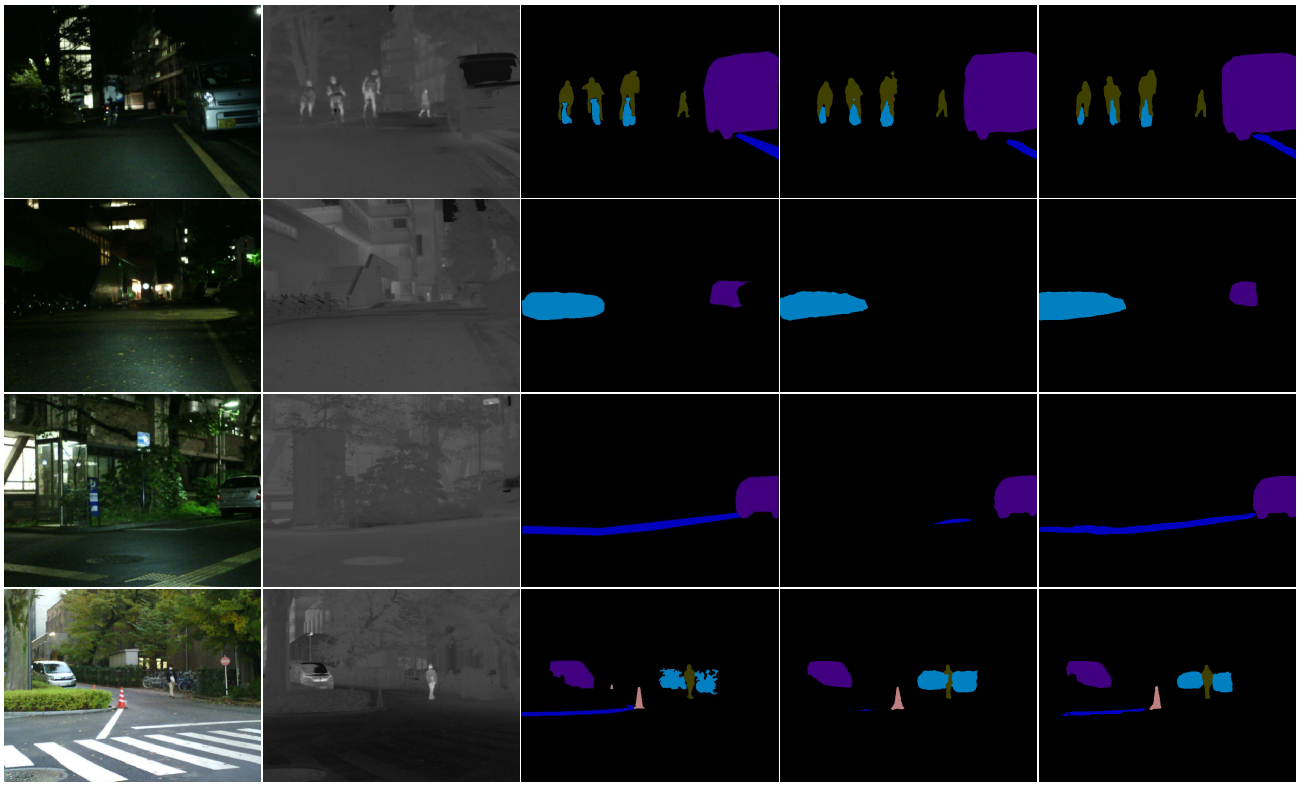
2.6. Qualitative Comparison Results

As shown in Figure 1, Figure 2, Figure 3 and Figure 4, we compare the downstream task performance of our method with CMNeXt [45] on four datasets: NYU Depth V2 [29] (RGB-D), MFNet [7] (RGB-T), MCubeS [16] (RGB-A-D), and DELIVER [45] (RGB-E). Our method demonstrates strong robustness under challenging conditions such as low-light and occlusion. For example, as shown in Figure 1, the second row illustrates accurate segmentation of the pillow under occlusion, while the fourth row shows stable performance under low-light noise. Similarly, in Figure 2, the second row shows precise segmentation of the purple vehicle in low-light conditions. In addition, our method produces smoother transitions and less noise along object boundaries. As shown in Figure 1 (third row), the transition between the cabinet and the wall is well preserved. Similar improvements can be observed in Figure 3 (third row) at the transition between water (purple) and rubber (light blue) materials, and in Figure 4 (second row) for the distinction between the road and grass.



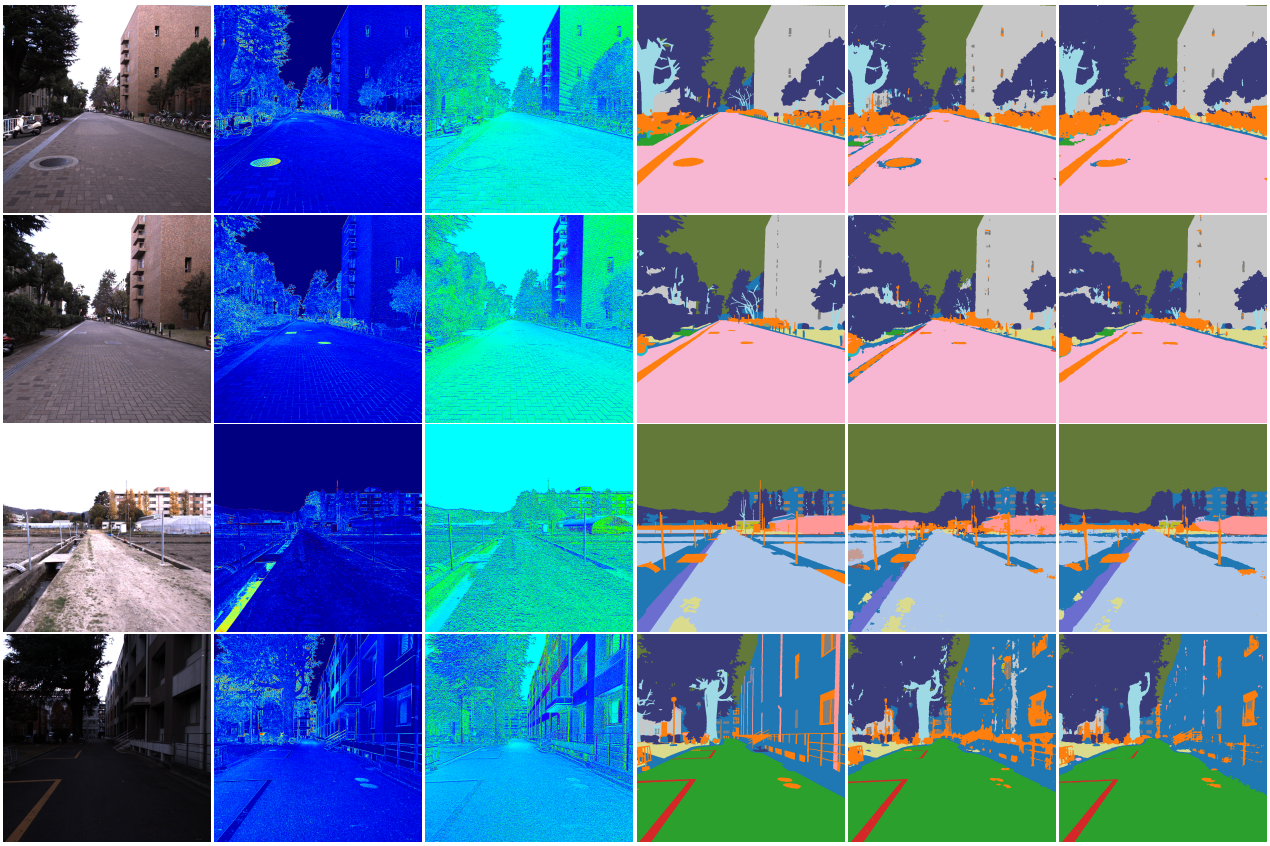
(a) RGB (b) Depth (c) GT (d) CMNeXt (e) Ours

Figure 1. Qualitative comparison results with full fine-tuning method CMNeXt [45] on NYU Depth V2.



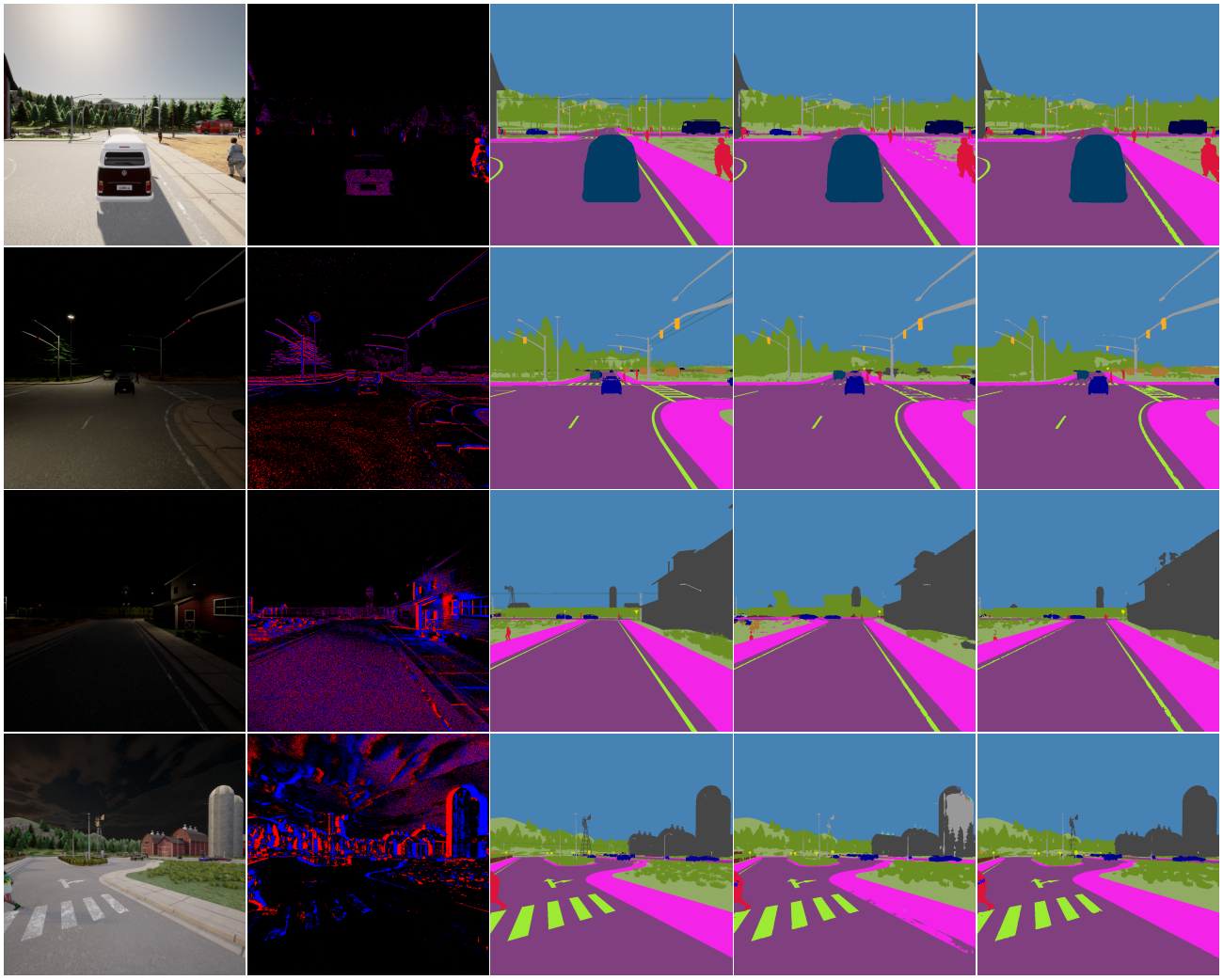
(a) RGB (b) Thermal (c) GT (d) CMNeXt (e) Ours

Figure 2. Qualitative comparison results with full fine-tuning method CMNeXt [45] on MFNet [7].



(a) RGB (b) DoLP (c) AoLP (d) GT (e) CMNeXt (f) Ours

Figure 3. Qualitative comparison results with full fine-tuning method CMNeXt [45] on MCubeS [16].



(a) RGB (b) Event (c) GT (d) CMNeXt (e) Ours

Figure 4. Qualitative comparison results with full fine-tuning method CMNeXt [45] on DELIVER [45].

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 2

[2] Nuo Chen, Chao Xiao, Yimian Dai, Shiman He, Miao Li, and Wei An. Event-based tiny object detection: A benchmark dataset and baseline. *arXiv preprint arXiv:2506.23575*, 2025. 2

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2

[5] Wen Dong, Haiyang Mei, Ziqi Wei, Ao Jin, Sen Qiu, Qiang Zhang, and Xin Yang. Exploiting polarized material cues for robust car detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1564–1572, 2024. 2

[6] FLIR. Free teledyne flir thermal dataset for algorithm training. *https://www.flir.com/oem/adas/ adas-dataset-form/*, 2018. 2

[7] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115, 2017. 1, 3, 5

[8] Lupiao Hu, Fasheng Wang, Fangmei Chen, Fuming Sun, and Haojie Li. Breaking alignment barriers: Tps-driven semantic correlation learning for alignment-free rgb-t salient object detection. *arXiv preprint arXiv:2512.21856*, 2025. 1

[9] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 2

[10] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021. 2

[11] Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify anything: Scaling indoor 3d object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22225–22233, 2025. 1

[12] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1750–1758, 2020. 2

[13] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. 2

[14] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 2

[15] Guoqiang Liang, Kanghao Chen, Hangyu Li, Yunfan Lu, and Lin Wang. Towards robust event-guided low-light image enhancement: a large-scale real-world event-image dataset and novel approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23–33, 2024. 2

[16] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 19800–19808, 2022. 1, 2, 3, 6

[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. *arXiv preprint arXiv:2503.06700*, 2025. 2, 3

[18] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022. 2

[19] Haoyue Liu, Shihan Peng, Lin Zhu, Yi Chang, Hanyu Zhou, and Luxin Yan. Seeing motion at nighttime with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25648–25658, 2024. 2

[20] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5811, 2022. 2

[21] Qiao Liu, Zhenyu He, Xin Li, and Yuan Zheng. Ptb-tir: A thermal infrared pedestrian tracking benchmark. *IEEE Transactions on Multimedia*, 22(3):666–675, 2019. 2

[22] Qiao Liu, Xin Li, Di Yuan, Chao Yang, Xiaojun Chang, and Zhenyu He. Lsotb-tir: A large-scale high-diversity thermal infrared single object tracking benchmark. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7): 9844–9857, 2023. 2

[23] Zhengyi Liu, Xinrui Wang, Xianyong Fang, Zhengzheng Tu, and Linbo Wang. Samsod: Rethinking sam optimization for rgb-t salient object detection. *IEEE Transactions on Multimedia*, 2026. 1

[24] Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. Glass segmentation using intensity and spectral polarization cues. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12622–12631, 2022. 2

[25] Yu Qiao, Bo Dong, Ao Jin, Yu Fu, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. Multi-view spectral polarization propagation for video glass segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23218–23228, 2023. 2

[26] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 2

[27] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008. 2

[28] 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 *International Conference on Robotics and Automation (ICRA)*, pages 9441–9447, 2020. 1, 2

[29] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760, 2012. 1, 3

[30] Kechen Song, Xiaotong Xue, Hongwei Wen, Yingying Ji, Yunhui Yan, and Qinggang Meng. Misaligned visible-thermal object detection: A drone-based benchmark and baseline. *IEEE Transactions on Intelligent Vehicles*, 2024. 2

[31] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 567–576, 2015. 1

[32] Chuanming Tang, Xiao Wang, Ju Huang, Bo Jiang, Lin Zhu, Shifeng Chen, Jianlin Zhang, Yaowei Wang, and Yonghong Tian. Revisiting color-event based tracking: A unified network, dataset, and metric. *Pattern Recognition*, page 112718, 2025. 2

[33] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 2

[34] Johan Vertens, Jannik Zürn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8461–8468, 2020. 2

[35] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE Transactions on Cybernetics*, 54(3): 1997–2010, 2023. 2

[36] Xiao Wang, Shiao Wang, Chuanming Tang, Lin Zhu, Bo Jiang, Yonghong Tian, and Jin Tang. Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19248–19257, 2024. 2

[37] Xiao Wang, Yu Jin, Wentao Wu, Wei Zhang, Lin Zhu, Bo Jiang, and Yonghong Tian. Object detection using event camera: A moe heat conduction based detector and a new benchmark dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29321–29330, 2025. 2

[38] Zheng Wu, Nathan Fuller, Diane Theriault, and Margrit Betke. A thermal infrared video benchmark for visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 201–208, 2014. 2

[39] Kaite Xiang, Kailun Yang, and Kaiwei Wang. Polarization-driven semantic segmentation via efficient attention-bridged fusion. *Optics Express*, 29(4):4802–4820, 2021. 2

[40] Xiuyu Xu, Minjie Wan, Junyu Ge, Hao Chen, Xingcheng Zhu, Xiaojie Zhang, Qian Chen, and Guohua Gu. Color-polarnet: residual dense network-based chromatic intensity-polarization imaging in low-light environment. *IEEE Transactions on Instrumentation and Measurement*, 71:1–10, 2022. 2

[41] Mingde Yao, Menglu Wang, King-Man Tam, Lingen Li, Tianfan Xue, and Jinwei Gu. Polarfree: Polarization-based reflection-free imaging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10890–10899, 2025. 2

[42] Minghao Ye, Xiao Liu, Yu Wang, Lu Liu, and Haoyao Chen. Rgb-thermal visual place recognition via vision foundation model. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4954–4960, 2025. 1

[43] Xinyi Ying, Chao Xiao, Wei An, Ruojing Li, Xu He, Boyang Li, Xu Cao, Zhaoxu Li, Yingqian Wang, Mingyuan Hu, et al. Visible-thermal tiny object detection: A benchmark dataset and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[44] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 1136–1147, 2023. 1

[45] 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, 3, 4, 5, 6, 7

[46] Yi Zhang, Wang Zeng, Sheng Jin, Chen Qian, Ping Luo, and Wentao Liu. When pedestrian detection meets multi-modal learning: Generalist model and benchmark dataset. In *European Conference on Computer Vision*, pages 430–448, 2024. 2

[47] Chu Zhou, Minggui Teng, Youwei Lyu, Si Li, Chao Xu, and Boxin Shi. Polarization-aware low-light image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3742–3750, 2023. 2