—Supplementary Material—
Multispectral Video Semantic Segmentation: A Benchmark Dataset and Baseline

Wei Ji[1] Jingjing Li[1,*] Cheng Bian[2] Zongwei Zhou[3]
Jiaying Zhao[2] Alan Yuille[3] Li Cheng[1]
[1]University of Alberta [2]ByteDance [3]Johns Hopkins University
{wji3, jingjin1, lcheng5}@ualberta.ca, zzhou82@jh.edu, ayuille1@jhu.edu
https://jiwei0921.github.io/Multispectral-Video-Semantic-Segmentation

Abstract

In this supplementary material, we first summarize the definitions of notations used in this paper in Sec. 1, and describe the detailed training setups in Sec. 2. In Sec. 3, we provide more details about the proposed MVSeg dataset. Then in Sec. 4, we provide more quantitative results, including the segmentation results on the validation set of MVSeg, the segmentation results of each class in the test set of MVSeg, and the efficiency results on the test of MVSeg. In Sec. 5, we show more qualitative visualizations on diverse scenarios of the MVSeg dataset, including daytime, nighttime with dim light, nighttime with overexposure, rainy, and snowy scenarios. These experimental results consistently demonstrate the superiority of our method to engage multispectral video data for semantic segmentation. Finally, we discuss potential future research directions in Sec. 6.

1. Notation and Definition

In Table 1, we summarize the notations and corresponding definitions used in this paper for better understanding.

2. Training Setups

Our code is implemented on the Pytorch platform and trained using two Nvidia RTX A6000 GPUs. Here we adopt various segmentation networks as the encoders (*e.g.,* DeepLabv3+ [1] with ResNet50 [10]). For the RGB stream, we initialize the network parameters using weights pre-trained on ImageNet [7]. For the thermal stream, we randomly initialize the network parameters, and generate 3-channel thermal images as inputs by repeating the 1-channel thermal images. Each image is uniformly resized to 320×480, and we perform random horizontal flipping and cropping to avoid potential over-fitting. Following [29], we use Adam optimizer with an initial learning rate of 2e-4, which is adaptively scheduled based on training loss: the learning rate is kept constant until the loss reaches a minimum plateau, after which we decrease it by 20%, until the learning rate reaches a minimum value of 1e-8. We set the batch size to 2. Following [2], τ is set to 0.1. λ is set to 0.001 empirically. We select 3 memory frames (*i.e.*, $L=3$), as discussed in ablation section. The network training involves two-stages: the first stage is backbone warming-up trained with only annotated query frames (150 epochs), and the second stage is main-training of MVNet trained with query and memory frames (200 epochs). In Table 2, we conduct comparison experiments to evaluate our two-stage training strategy and one-stage training with main-training only. It is shown that integrating our backbone warming-up stage improves the mIoU score by 1.53%. This is because the backbone warming-up process enables the model to infer more meaningful semantic masks for memory frames at the early training stage, thereby providing representative prototypical features for training the entire MVNet.

3. More Dataset Details

Dataset preparation: We gathered RGB-thermal videos from multiple sources in related works, including OSU [5], INO [14], RGBT234 [20], and KAIST [13]. It is known that the multispectral datasets suffer from the misalignment issue, even in the KAIST dataset which has improved alignment by a beam splitter. To further address this issue, we implemented several proactive measures. Specifically, we performed a visualization screening process by overlaying thermal heat maps onto paired RGB images, making it easier for inspectors to verify alignment. This process resulted in a filtering rate of 77% (180K/233K), effectively removing relatively low-quality image pairs and ensuring the quality of our MVSeg dataset to a large extent. Additionally, for videos with minor misalignment near the edges, we cropped and retained the central regions to minimize misalignment issues. Ultimately, we selected 738 high-quality video shots (averaging 5 seconds each) to comprise our MVSeg dataset.

Table 1. Summary of the notations and corresponding definitions used in this paper.

Notation	Definition
t	The time subscript for query (current) frame.
L	The number of past frames used in memory.
U	The set of time scripts of query and memory frames; $U = \{t-L, \ldots, t-1, t\}$.
d	The time subscript of a certain frame in the set of U.
$\mathcal{M}$	The set of modality types; $\mathcal{M} = \{R, T, F\}$, representing RGB, thermal, and fused types, respectively.
m	A specific modality type in the set of $\mathcal{M}$.
I_d^m	The multispectral video input, in which $d \in U, m \in \{R, T\}$.
$\mathbf{f}_d^m \in \mathbb{R}^{H \times W \times D}$	The extracted multispectral video features, where $H \times W$ represents the spatial size, D is the channel dimension, and $m \in \mathcal{M}$, $d \in U$.
$\mathcal{C}$	The set of semantic categories.
c, $\bar{c}$	A specific semantic category in the set of $\mathcal{C}$.
$\lvert\mathcal{C}\rvert$	The number of semantic categories.
$\mathbf{p}^m$	The prototypical memory feature on each modality, where $\{\mathbf{p}^m \in \mathbb{R}^{L\lvert\mathcal{C}\rvert \times D}\}_{m \in \mathcal{M}}$.
$\mathbf{w}^m$	The intermediate weighting maps in MVFuse module.
$\mathcal{P}, \mathcal{N}$	The positive set and negative set in MVRegulator loss.
$\mathbf{F}_d^m \in \mathbb{R}^{H \times W \times D}$	The memory-augmented multispectral video features building upon $\mathbf{f}_d^m$.

Table 2. Ablation analysis of training schemes, where the models trained through different training strategies.

Settings	mIoU(%)
Main-training only	52.99
Main-training with backbone warming-up (Ours)	54.52

Dataset composition and split: The composition and split of the MVSeg dataset are summarized in Table 3, where we pay attention to provide a reasonable distribution of video sources across sets without obvious domain shift.

Table 3. Detailed composition and split of the MVSeg dataset.

	Overall		Train		Val		Test	
	#Vids	#GTs	#Vids	#GTs	#Vids	#GTs	#Vids	#GTs
INO [14]	160	811	107	549	19	92	34	170
KAIST [13]	332	1585	204	1013	32	142	96	430
OSU [5]	8	40	3	20	2	6	3	14
RGBT234 [20]	238	1109	138	659	31	138	69	312
Total	738	3545	452	2241	84	378	202	926

4. Additional Quantitative Results

In this section, we provide more quantitative results of our MVNet and related approaches. Table 4 presents the segmentation results on the validation set of MVSeg. It is observed that, compared to the image-based segmentation models (*i.e.*, FCN [26], PSPNet [39], and DeepLabv3+ [1]), our MVNet variants consistently bring substantial performance gains of 2.73%, 2.85% and 2.53%, respectively. This proves the superiority of our MVNet in incorporating multispectral video data to improve semantic segmentation accuracy. In Table 5, we also provide detailed results on each class of the MVSeg test set for reference. In Table 6, we show more results of related MSS/VSS methods, including CNN-based and transformer-based networks. We also apply our method to the transformer-based image segmentation network, SegFormer (MiT-B1&-B2) [35], to provide a more thorough validation. The model parameters, GPU memory usage during training, and inference time (ms) per frame are also included. These results consistently verify the benefits of incorporating multispectral temporal contexts for semantic segmentation as well as the superiority of our approach.

5. Additional Qualitative Results

In this section, we provide more qualitative results of our MVNet and related approaches. We first visualize more segmentation results of diverse scenarios in the test set of MVSeg, including common daytime scenes, nighttime with dim light, nighttime with overexposure, rainy, and snowy scenarios, as shown in Fig. 2 to Fig. 6. For each video example, we show three representative video frames of both RGB and thermal modes and the segmentation results of various models. We highlight the details with the yellow boxes. Obviously, the results from our MVNet model are more accurate compared to the competing methods. We owe this to the superiority of our method in engaging the advantages of complementary multispectral and temporal contexts. In addition, these visualizations of different scenarios also showcase the diversity of our MVSeg dataset, which is expected to provide a sufficiently realistic benchmark in this field.

To give an intuitive view, we also generate a video demo as in our github website. It is shown that the incorporation of multispectral video information can help produce more reliable and robust segmentation results when facing diverse lighting conditions. In addition, we find that the accuracy of MVSS model still has a large room for improvement, which

Table 4. Quantitative evaluation on the validation set of MVSeg dataset. The notation † and ‡ are used to mark the VSS and MSS models, respectively.

Method	Backbone	mIoU(%)
CCNet [12]	ResNet-50	50.90
OCRNet [38]	ResNet-50	51.99
STM† [28]	ResNet-50	52.55
LMANet† [29]	ResNet-50	52.69
MFNet‡ [9]	Mini-inception	51.17
RTFNet‡ [33]	ResNet-152	52.65
EGFNet‡ [40]	ResNet-152	52.73
FCN [26]	ResNet-50	50.32
MVNet$_{FCN}$	ResNet-50	53.05 (+2.73)
PSPNet [39]	ResNet-50	50.58
MVNet$_{PSPNet}$	ResNet-50	53.43 (+2.85)
DeepLabv3+ [1]	ResNet-50	50.77
MVNet$_{DeepLabv3+}$	ResNet-50	53.30 (+2.53)

is deserved to be explored in future works.

RGB Frame Optical Flow RGB Frame Optical Flow

Figure 1. Visualizations of the estimated optical flow on MVSeg benchmark, where we adopt the widely-used RAFT model [34] following the standard protocol of [27].

6. Discussion

Here we discuss five potential research problems on the proposed multispectral video semantic segmentation. Meanwhile, some feasible solutions are given for reference.

i) *Accuracy*: The research of MVSS is still in its initial stage. By drawing ideas from the well-studied semantic segmentation of RGB images, the accuracy of MVSS model can be further advanced. For example, we may integrate the multi-scale learning technique [1, 30, 36] into cross-spectral and cross-frame fusion to improve the contextual representability of MVSS models; we may also introduce extra edge signals [16, 23] to help the model capture object boundary details. In addition, it is worth exploring more advanced fusion techniques [17, 22, 40] to promote sufficient interactions among multimodal information.

ii) *Efficiency*: Although the engagement of multispectral videos brings significant improvement, it introduces additional model parameters. More lightweight models need to be explored to improve efficiency. For example, we can adopt more lightweight operations such as depthwise separable convolution [4] or neural architecture search techniques [24], to advance feature extractors. On the other hand, we may explore knowledge distillation scheme [11] to transfer thermal knowledge to RGB stream, which can avoid the heavy overhead of thermal encoder.

iii) *Evaluation Metrics*. Due to the challenging scenes in MVSeg benchmark, the popular TC metric [25] that evaluates temporal consistency based on optical flow warping may not correctly reflect the performance of MVSS models. As illustrated in Fig. 1, the estimated optical flows of complex nighttime scenes is not meaningful, which cannot well represent the motions of objects in the scene, *e.g.,* the less-visible driving cars in dim night. Thus, how to design suitable metrics for MVSS is still an open issue.

iv) *Weak Supervision.* The acquirement of per-pixel semantic labels is laborious and time-consuming. Thus, training MVSS models with weak supervisions [3, 15, 19, 21] (*e.g.*, image-level category, bounding box) is an appealing future direction, which can avoid heavy annotation costs.

v) *Instance-level Extension.* As described in [37], video instance segmentation is more crucial than object-level semantic segmentation for practical applications [8, 18]. Thus, the research on multispectral video instance segmentation is an essential future direction. Moving forward, we will prioritize our efforts to further explore this area.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 1, 2, 3, 4, 5, 6, 7

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 1

[3] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *CVPR*, pages 2617–2626, 2022. 3

[4] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017. 3

[5] James W Davis and Vinay Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(2-3):162–182, 2007. 1, 2

[6] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In *IROS*, pages 4467–4473, 2021. 4

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[8] Simon Graham, Mostafa Jahanifar, Ayesha Azam, Mohammed Nimir, Yee-Wah Tsang, Katherine Dodd, Emily Hero, Harvir Sahota, Atisha Tank, Ksenija Benes, et al.

Table 5. The quantitative results on each class of the MVSeg benchmark test set. 'Deep3+' is the shorthand for DeepLabv3+ [1].

*	CCNet [12]	OCRNet [38]	STM [28]	LMANet [29]	MFNet [9]	RTFNet [33]	EGFNet [40]	FCN [26]	**MVNet** (*FCN*)	PSPNet [39]	**MVNet** (*PSPNet*)	Deep3+ [1]	**MVNet** (*Deep3+*)
Background	38.40	40.39	39.13	36.63	36.21	34.31	34.48	36.00	35.38	34.37	39.79	35.09	39.79
Car	79.75	81.82	79.59	81.35	80.54	81.43	80.77	78.64	81.39	80.74	81.45	80.66	80.98
Bus	36.82	37.53	34.59	36.48	40.03	35.66	41.61	36.42	37.40	39.69	38.95	36.87	31.50
Motorcycle	29.78	36.02	32.94	36.57	31.59	32.34	33.41	30.92	33.78	32.66	36.42	37.40	35.39
Bicycle	60.04	62.60	60.17	57.67	53.89	55.12	61.41	52.56	60.70	54.71	59.59	55.24	65.34
Pedestrian	55.00	55.59	54.05	58.95	58.04	61.76	61.75	52.36	61.63	57.39	59.18	54.46	58.30
Motorcyclist	4.1	4.43	3.33	12.23	6.22	13.83	11.88	8.8	10.00	4.32	8.12	6.32	8.22
Bicyclist	27.21	25.48	25.52	32.08	26.06	30.64	31.75	23.27	33.27	29.28	30.10	33.43	35.95
Cart	49.22	58.00	54.17	46.41	49.73	50.60	54.17	48.93	50.97	57.25	56.07	50.42	58.40
Bench	59.64	53.18	58.05	51.95	51.54	55.57	54.22	59.72	54.31	54.46	56.90	33.94	50.39
Umbrella	40.72	42.97	45.68	34.32	36.07	45.95	44.75	33.83	43.19	37.70	47.67	43.83	45.05
Box	36.66	37.87	38.98	39.69	40.11	37.83	38.79	37.70	41.53	37.81	39.00	37.83	41.85
Pole	51.12	50.33	54.12	49.01	46.31	44.39	45.57	49.19	51.44	45.62	55.36	51.36	53.33
Street Lamp	49.50	56.84	61.41	55.45	53.50	58.11	50.74	54.91	62.13	57.19	59.98	53.63	59.79
Traffic Light	35.68	28.54	40.93	36.80	36.86	37.71	35.67	37.27	43.37	35.53	39.24	39.10	42.30
Traffic Sign	35.91	37.24	43.25	41.12	37.38	36.53	44.11	38.10	40.84	40.18	42.58	39.29	41.22
Car Stop	35.63	37.25	32.36	33.34	34.37	33.91	35.50	35.26	34.95	29.77	34.53	32.99	35.88
Color Cone	22.56	20.12	22.07	24.52	18.19	15.84	21.02	4.8	13.87	18.47	23.83	15.55	24.42
Sky	84.31	84.29	88.25	87.40	90.53	90.88	90.27	91.10	87.84	77.93	87.43	84.55	88.02
Ground	86.43	88.04	78.97	89.16	86.12	86.75	85.79	87.26	88.51	87.05	85.62	89.32	88.19
Road	90.45	89.79	88.52	90.99	90.85	90.73	90.70	90.27	91.24	90.32	91.10	91.32	91.23
Sidewalk	54.88	56.13	51.06	58.90	57.75	57.07	58.15	54.07	58.01	58.43	57.28	57.58	57.65
Curb	49.13	45.94	49.25	48.93	46.79	49.35	47.81	44.55	48.44	46.54	50.88	48.53	48.41
Terrain	77.37	76.69	75.24	76.27	77.94	78.67	79.06	75.63	78.70	76.69	75.71	77.75	78.03
Vegetation	78.37	78.31	78.41	77.74	79.09	80.24	79.92	79.57	80.07	75.81	79.34	78.52	80.14
Building	75.62	76.54	75.16	77.06	76.62	76.93	76.19	76.40	78.45	75.86	77.18	76.27	77.68
mean IoU (%)	51.70	52.38	52.51	52.73	51.63	52.77	53.44	50.67	53.90	51.38	54.36	51.59	**54.52**

Table 6. More quantitative evaluation on the test set of the MVSeg dataset. The notation † and ‡ stand for the VSS and MSS models, respectively. $\star$ denotes transformer-based models that have input image with size 480×480.

Methods	Backbone	#Param(M)	#Mem(G)	Times(ms)	mIoU(%)
CCNet [12]	ResNet-50	52.3	4.9	10.3	51.70
OCRNet [38]	ResNet-50	43.6	4.8	9.8	52.38
STM† [28]	ResNet-50	44.1	21.9	11.2	52.51
LMANet† [29]	ResNet-50	44.1	10.4	9.7	52.73
CFFM†* [32]	MiT-B1	15.5	13.3	14.1	52.83
MFNet‡ [9]	Mini-inception	0.72	2.3	5.3	51.63
PSTNet‡ [31]	ResNet-18	29.0	3.2	3.5	51.78
RTFNet‡ [33]	ResNet-152	254.5	8.6	47.7	52.77
FEANet‡ [6]	ResNet-152	255.2	8.7	34.3	53.19
EGFNet‡ [40]	ResNet-152	123.2	6.3	75.8	53.44
DeepLabv3+ [1]	ResNet-50	41.6	4.6	8.1	51.59
Ours (DeepLabv3+)	ResNet-50	88.4	18.8	18.4	54.52
SegFormer* [35]	MiT-B1	13.8	4.0	7.9	51.11
Ours (SegFormer)*	MiT-B1	33.7	14.1	17.6	54.25
SegFormer* [35]	MiT-B2	24.8	4.6	13.7	53.07
Ours (SegFormer)*	MiT-B2	56.1	18.6	29.8	55.22

Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification. In *ICCV*, pages 684–693, 2021. 3

[9] 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 *IROS*, pages 5108–5115, 2017. 3, 4

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[11] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 3

[12] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, pages 603–

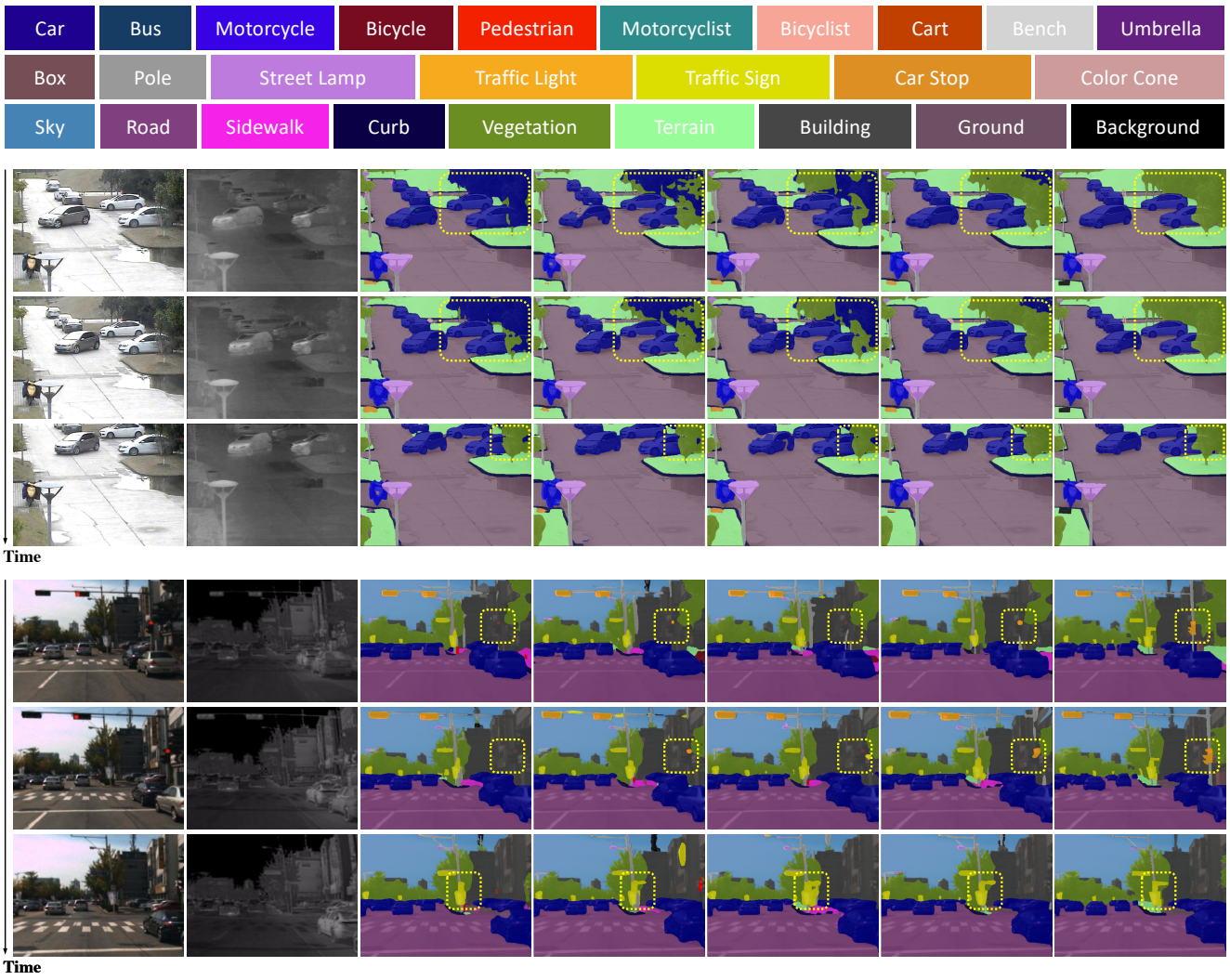

Figure 2. Qualitative semantic segmentation results on the common ***daytime*** scenarios. From left to right: RGB frames, thermal infrared (TIR) frames, results of DeepLabv3+ [1], LMANet [29], EGFNet [40] as well as our proposed MVNet, and ground truth of multispectral video semantic segmentation. We highlight the improved details with the yellow boxes. Best viewed in color and zoom in.

612, 2019. 3, 4

[13] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, pages 1037–1045, 2015. 1, 2

[14] INO. Video analytics dataset. https://www.ino.ca/en/technologies/video-analytics-dataset/, 2012. 1, 2

[15] Wei Ji, Jingjing Li, Qi Bi, Chuan Guo, Jie Liu, and Li Cheng. Promoting saliency from depth: Deep unsupervised rgb-d saliency detection. *ICLR*, 2022. 3

[16] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate rgb-d salient object detection via collaborative learning. In *ECCV*, pages 52–69. Springer, 2020. 3

[17] 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 Transactions on Image Processing*, 31:2321–2336, 2022. 3

[18] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *CVPR*, pages 12341–12351, 2021. 3

[19] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, pages 876–885, 2017. 3

[20] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. 1, 2

[21] Jingjing Li, Wei Ji, Qi Bi, Cheng Yan, Miao Zhang, Yongri Piao, Huchuan Lu, et al. Joint semantic mining for

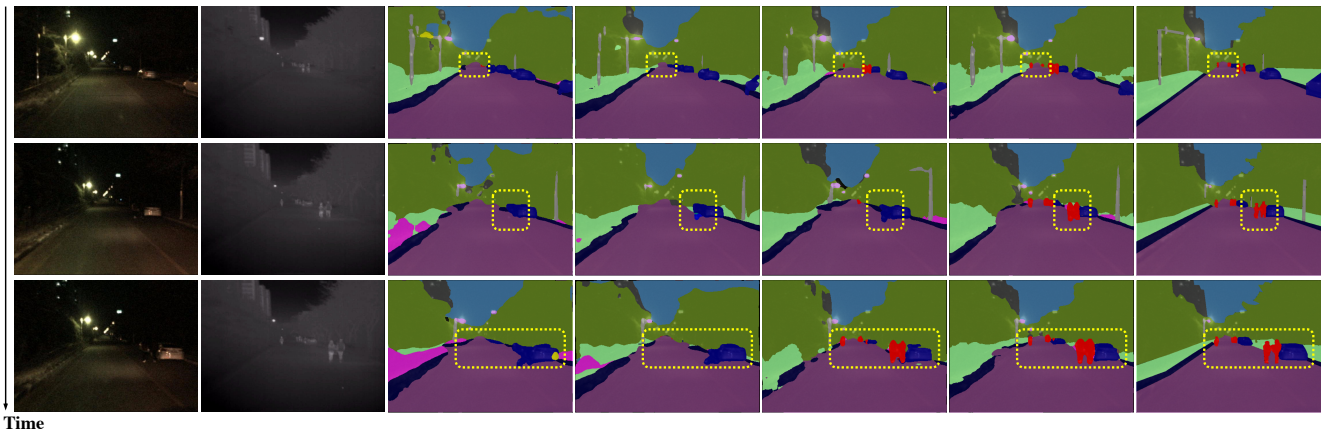

Figure 3. Qualitative semantic segmentation results on the ***nighttime & dim light*** scenario. From left to right: RGB frames, thermal infrared (TIR) frames, results of DeepLabv3+ [1], LMANet [29], EGFNet [40] as well as our proposed MVNet, and ground truth of multispectral video semantic segmentation. We highlight the improved details with the yellow boxes. Best viewed in color and zoom in.

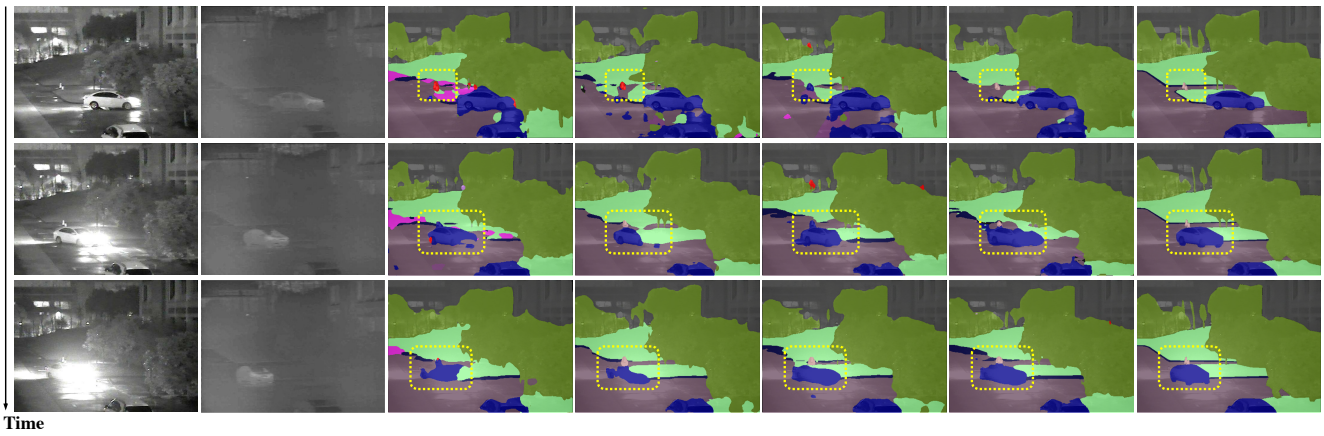

Figure 4. Qualitative semantic segmentation results on the ***nighttime & overexposure*** scenario. From left to right: RGB frames, thermal infrared (TIR) frames, results of DeepLabv3+ [1], LMANet [29], EGFNet [40] as well as our proposed MVNet, and ground truth of multispectral video semantic segmentation. We highlight the improved details with the yellow boxes. Best viewed in color and zoom in.

weakly supervised rgb-d salient object detection. *NeurIPS*, 34:11945–11959, 2021. 3

[22] Jingjing Li, Wei Ji, Miao Zhang, Yongri Piao, Huchuan Lu, and Li Cheng. Delving into calibrated depth for accurate rgb-d salient object detection. *International Journal of Computer Vision*, pages 1–22, 2022. 3

[23] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *ECCV*, pages 435–452, 2020. 3

[24] Yingwei Li, Xiaojie Jin, Jieru Mei, Xiaochen Lian, Linjie Yang, Cihang Xie, Qihang Yu, Yuyin Zhou, Song Bai, and Alan L. Yuille. Neural architecture search for lightweight non-local networks. In *CVPR*, June 2020. 3

[25] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *European Conference on Computer Vision*, pages 352–368. Springer, 2020. 3

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2, 3, 4

[27] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *CVPR*, pages 4133–4143, 2021. 3

[28] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, pages 9226–9235, 2019. 3, 4

[29] Matthieu Paul, Martin Danelljan, Luc Van Gool, and Radu Timofte. Local memory attention for fast video semantic segmentation. In *IROS*, pages 1102–1109, 2021. 1, 3, 4, 5, 6, 7

[30] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019. 3

[31] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-

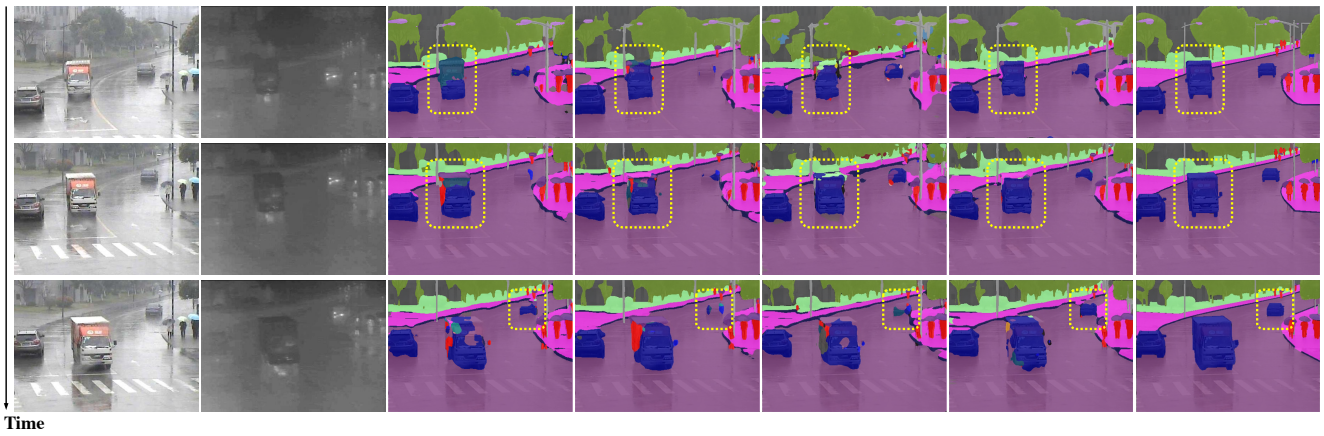

Figure 5. Qualitative semantic segmentation results on the ***rainy*** scenario. From left to right: RGB frames, thermal infrared (TIR) frames, results of DeepLabv3+ [1], LMANet [29], EGFNet [40] as well as our proposed MVNet, and ground truth of multispectral video semantic segmentation. We highlight the improved details with the yellow boxes. Best viewed in color and zoom in.

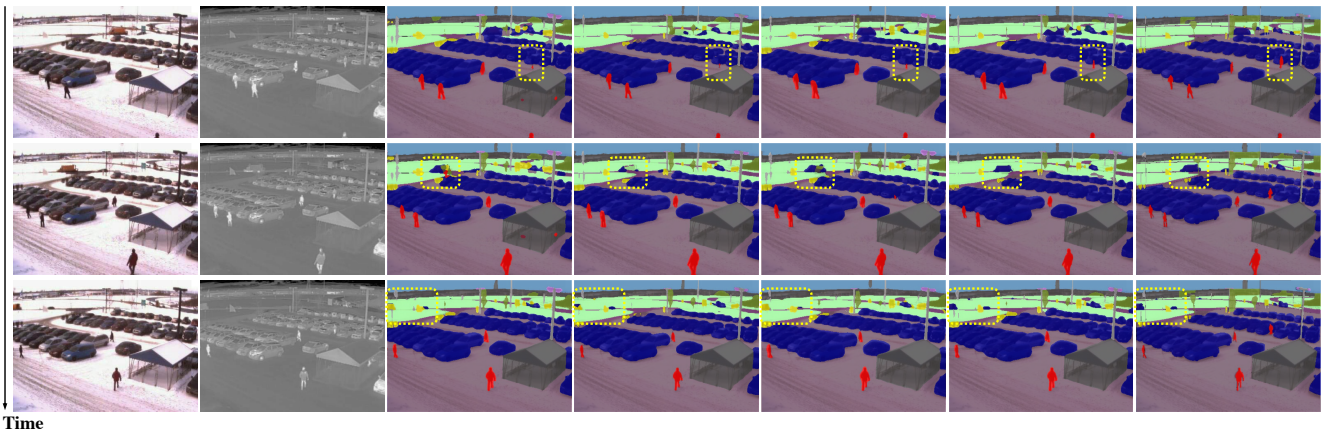

Figure 6. Qualitative semantic segmentation results on the ***snowy*** scenario. From left to right: RGB frames, thermal infrared (TIR) frames, results of DeepLabv3+ [1], LMANet [29], EGFNet [40] as well as our proposed MVNet, and ground truth of multispectral video semantic segmentation. We highlight the improved details with the yellow boxes. Best viewed in color and zoom in.

thermal calibration, dataset and segmentation network. In *ICRA*, pages 9441–9447, 2020. 4

[32] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *CVPR*, pages 3126–3137, 2022. 4

[33] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019. 3, 4

[34] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2020. 3

[35] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34, 2021. 2, 4

[36] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal attention for long-range interactions in vision transformers. In *NeurIPS*, volume 34, pages 30008–30022, 2021. 3

[37] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, pages 5188–5197, 2019. 3

[38] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, pages 173–190, 2020. 3, 4

[39] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 2, 3, 4

[40] Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for rgb thermal scene parsing. In *AAAI*, 2022. 3, 4, 5, 6, 7