Delivering Arbitrary-Modal Semantic Segmentation

(Supplementary Material)

A. DELIVER Dataset

A.1. Detailed settings in data collection

Depth2Frames. The depth camera straightforwardly outputs a grayscale depth map (*i.e.* 0–255 scales), which will cause discontinuity and quantization errors in distance measurements. Therefore, we convert the original depth image to the depth frame using a logarithmic scale, leading to milimetric granularity and better precision at close ranges.
Event2Frames. The positive- and negative event threshold of the event camera are both set to 0.3. We record raw event point cloud between two adjacent frames and convert the last occurring event among all pixels into an event frame, where blue indicates positive and red indicates negative.
LiDAR2Frames. We transform the LiDAR point cloud to the image coordinate system, so as to obtain an image-like representation of LiDAR data. The Field-of-View (FoV) of the front camera is $91°$ and the image resolution is $H{\times}W{=}1042{\times}1042$. The origin is $(u_0, v_0){=}(H/2, W/2)$. The focal length (f_x, f_y) is calculated as:

$$f_x{=}H/(2{\times}tan(FoV{\times}\pi/360)), \quad (1)$$
$$f_y{=}W/(2{\times}tan(FoV{\times}\pi/360)). \quad (2)$$

To project 3D points to 2D image coordinate, we have:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \mathbf{0}_{3\times1}^T & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \quad (3)$$

where (X,Y,Z) is the LiDAR point, (u, v) is the 2D image pixel, and the rotation ($\boldsymbol{R}$) and the translation ($\boldsymbol{t}$) matrices are set as the unit matrix in the CARLA simulator [2].

A.2. Dataset structure

DELIVER contains Depth, LiDAR, Event, and RGB modalities. As shown in Fig. 1, four adverse road scene conditions of *rainy*, *sunny*, *foggy*, and *night* are included in our dataset. There are five sensor failure cases including Motion Blur (**MB**), Over-Exposure (**OE**), Under-Exposure (**UE**), LiDAR-Jitter (**LJ**), and Event Low-resolution (**EL**) to verify that the performance of model is robust and stable in the presence of sensor failures. The sensors are mounted at different locations on the ego car to provide multiple views including *front*, *rear*, *left*, *right*, *up*, and *down*. Each sample is annotated with semantic and instance labels. In this work, we focus on the front-view semantic segmentation.

The 25 semantic classes in DELIVER dataset are: *Building, Fence, Other, Pedestrian, Pole, RoadLine, Road, SideWalk, Vegetation, Cars, Wall, TrafficSign, Sky, Ground, Bridge, RailTrack, GroundRail, TrafficLight, Static, Dynamic, Water, Terrain, TwoWheeler, Bus, Truck*.

A.3. Dataset statistics

We present statistics of the DELIVER dataset in Table 1. We discuss data partitioning in two groups, one according to the conditions and the other according to the sensor failures. Note that, the two groups are mutually inclusive. The five cases from the second group are included in each of five conditions from the first group. For example, cases of **MB**, **OE**, **UE**, **LJ**, and **EL** are included in *cloudy*, *foggy*, *night*, *rainy*, and *sunny* conditions, but with different samples. To investigate the robustness under sensor failures, we collect 1199, 400, 398, 398, and 409 frames on respective cases.

A.4. Dataset comparison

As shown in Table 2, we compare several datasets with adverse conditions and cases. All the datasets cover the whole daytime. The real-scene datasets, *e.g.*, WildDash [12] and Waymo [9], capture data by using only one or a few sensors, which results a lack of data diversity. In contrast, our DELIVER dataset has four different modalities, including *RGB*, *Depth*, *Event* and *LiDAR*, which enables the multimodal semantic segmentation task to involve up to 4 modalities. Compared to previous synthetic datasets, *e.g.*, SELMA [10], SynWoodScape [7], SynPASS [13], our DELIVER additionally includes 5 types of sensor failure. Each sample has semantic and instance annotations, so semantic, instance and panoptic segmentation tasks can be conducted on our DELIVER dataset.

B. Implementation Details

We conduct our experiments with PyTorch $1.9.0$. All models are trained on a node with 4 A100 GPUs. Below we describe the specific implementation details for six datasets.
Data representation. For depth images, we follow SA-Gate [1] and CMX [6] to preprocess the one-channel depth images to HHA-encoded representations [4], where HHA includes horizontal disparity, height above ground, and norm angle. The 3D LiDAR and Event data of DELIVER dataset are transformed to the aforementioned frame format. Then, both LiDAR- and Event-based data are preprocessed as 2D range views [15] and 3-channel representations [14], respectively.

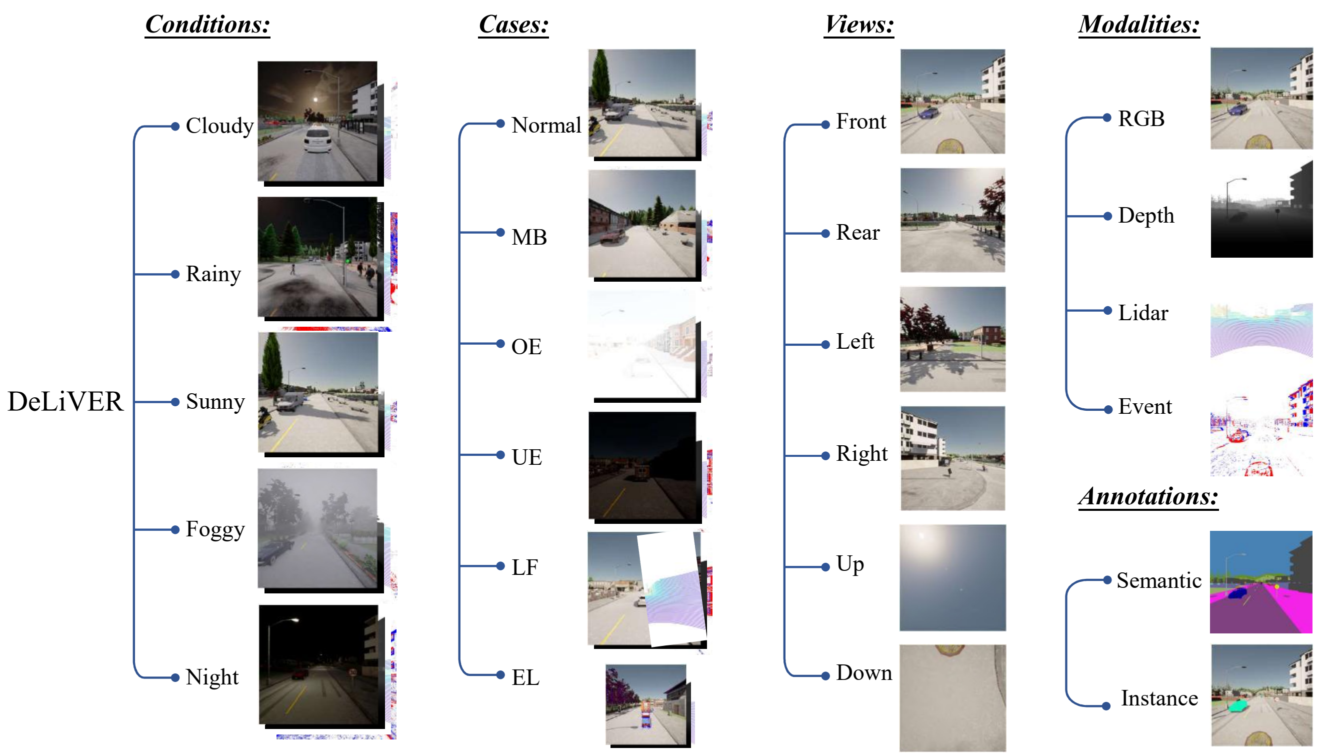

Figure 1. Data structure of the DELIVER dataset. The columns from left to right are respective conditions, cases, multiple views, modalities and annotations. **MB**: Motion Blur; **OE**: Over-Exposure; **UE**: Under-Exposure; **LJ**: LiDAR-Jitter; and **EL**: Event Low-resolution.

Table 1. Data statistic of DeLiVER dataset. It includes four adverse conditions (*cloudy*, *foggy*, *rainy*, and *night*), and each condition has five failure cases (**MB**: Motion Blur; **OE**: Over-Exposure; **UE**: Under-Exposure; **LJ**: LiDAR-Jitter; and **EL**: Event Low-resolution).

Split	**Cloudy**	**Foggy**	**Night**	**Rainy**	**Sunny**	**Total**	**Normal**	**MB**	**OE**	**UE**	**LJ**	**EL**	**Total**
Train	794	795	797	799	798	3983	2585	600	200	199	199	200	3983
Val	398	400	410	398	399	2005	1298	299	100	99	100	109	2005
Test	379	379	379	380	380	1897	1198	300	100	100	99	100	1897
Front-view	1571	1574	1586	1577	1577	7885	5081	1199	400	398	398	409	7885
All six views	9426	9444	9516	9462	9462	47310	30486	7194	2400	2388	2388	2454	47310

Table 2. Comparison between multimodal datasets. D:Day; S:Sunset; N:Night; *:random; Sem.:Semantic; Ins.:Instance.

Dataset	Type	Sensors				Sensor	RGB	Diversity			Classes	Labels	
		Camera	Depth	Event	LiDAR	Failures	Failures	Weathers	Daytime	Views		Sem.	Ins.
WildDash [12]	Real	1	0	0	0	0	15	*	*	*	19	✓	✓
Waymo [9]	Real	5	0	0	5	0	0	2	DN	5	28	✓	✓
SELMA [10]	Synthetic	7	7	0	3	0	6	9	DSN	7	19	✓	×
SynWoodScape [7]	Synthetic	5	5	5	1	0	0	4	DS	5	25	✓	✓
SynPASS [13]	Synthetic	6	0	0	0	0	0	4	DN	1	22	✓	×
DeLiVER (ours)	Synthetic	6	6	6	1	5	3	4	DN	6	25	✓	✓

DELIVER dataset. We train our models for 200 epochs on the DELIVER dataset. The batch size is 2 on each of four GPUs. The resolution of all modalities is set as 1024×1024 for training and inference. In the Event Low-resolution cases, the Event-based images with the original size of 260×260 are upsampled to 1024×1024. During evaluation, we only apply the single-scale test strategy. The backbone of CMNeXt is based on MiT-B2 [11]. To verify the effectiveness of our method under convolutional networks, the CNN-based SegNeXt-Base [3] is selected as the backbone, when compared to the MiT-B2 one.

KITTI-360 dataset. As there are more than $49K$ training data on KITTI-360 dataset, the models are trained for 40 epochs. The image resolution is set as 1408×376 and the

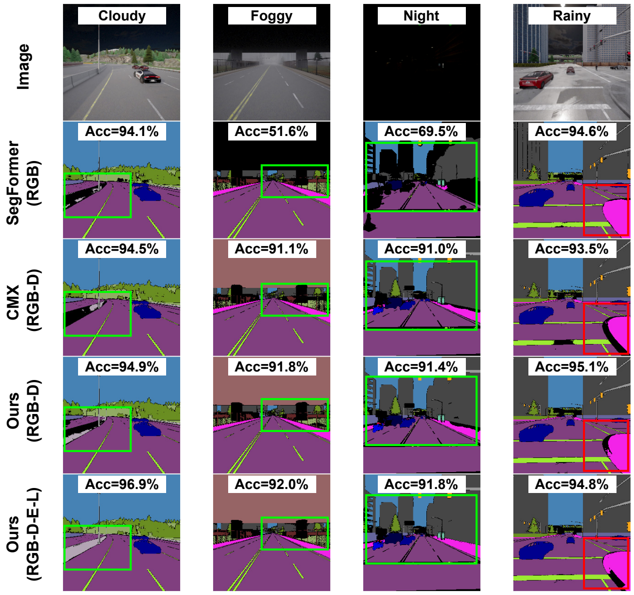

Figure 2. More visualization results on DeLiVER dataset. From left to right are the respective *cloudy*, *foggy*, *night* and *rainy* scene.

batch size is 4 on each of four GPUs. The backbone of CMNeXt is based on MiT-B2 [11].

NYU Depth V2 dataset. Following CMX [6], the number of training epochs is set as 500 for a fair comparison. The resolution of RGB and Depth images is set as 640×480. The training batch size is 4 on each of four GPUs. The backbone of CMNeXt is based on MiT-B4 [11]. We apply the multi-scale flip test strategy for a fair comparison.

MFNet dataset. We train our CMNeXt models with the MiT-B4 backbone for 500 epochs on the MFNet dataset. The resolution of RGB and Thermal images is set as 640×480 and the batch size is 4 on each of four GPUs. We apply the multi-scale flip test strategy for a fair comparison.

MCubeS dataset. To compare with MCubeSNet [5], we build CMNeXt with MiT-B2 and train the model for 500 epochs. Following MCubeSNet [5], the image size is set as 512×512 during training and 1024×1024 during evaluation. The batch size is set as 4 on each of four GPUs.

UrbanLF dataset. To perform comparison with the OCR-LF model [8], we build CMNeXt with MiT-B4. The image size on the real and synthetic sets is 640×480. The angular resolution of 81 sub-aperture images of the UrbanLF dataset is 9×9. To conduct arbitrary-modal segmentation, the center-aperture image is selected as the primary modality, while the other apertures are as additional modalities. We sample respective 8, 33, and 80 light field images as the supplementary modalities, *i.e.*, LF8, LF33, and LF80 for short. The 8 images are from the center horizontal direction, while the 33 images are from the four directions of horizontal, vertical, $\frac{1}{4}\pi$, and $\frac{3}{4}\pi$, following UrbanLF [8].

C. More visualizations on DeLiVER

As shown in Fig. 2, in the four adverse weather conditions, RGB-D fusion-based methods greatly improve the performance, particularly for distant elements in *foggy* and *nighttime* scenes. Our RGB-D solution is more accurate than CMX (RGB-D), and the full quad-modal RGB-D-E-L CMNeXt model further enhances the segmentation. A failure case is shown on the right column (*i.e.*, the *rainy* scene) of Fig. 2, in which the RGB-only model has a better segmentation on the *sidewalk* class. However, our quad-modal CMNeXt has a higher accuracy score with 94.8%.

D. Acknowledgments

This work was supported in part by Helmholtz Association of German Research Centers, in part by the Federal Ministry of Labor and Social Affairs (BMAS) through the AccessibleMaps project under Grant 01KM151112, in part by the University of Excellence through the “KIT Future Fields” project, and in part by Hangzhou SurImage Technology Company Ltd. This work was partially performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research.

References

[1] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In *ECCV*, 2020. 1

[2] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *CoRL*, 2017. 1

[3] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. SegNeXt: Rethinking convolutional attention design for semantic segmentation. In *NeurIPS*, 2022. 2

[4] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *ECCV*, 2014. 1

[5] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *CVPR*, 2022. 3

[6] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838*, 2022. 1, 3

[7] Ahmed Rida Sekkat, Yohan Dupuis, Varun Ravi Kumar, Hazem Rashed, Senthil Yogamani, Pascal Vasseur, and Paul Honeine. SynWoodScape: Synthetic surround-view fisheye camera dataset for autonomous driving. *RA-L*, 2022. 1, 2

[8] Hao Sheng, Ruixuan Cong, Da Yang, Rongshan Chen, Sizhe Wang, and Zhenglong Cui. UrbanLF: A comprehensive light field dataset for semantic segmentation of urban scenes. *TCSVT*, 2022. 3

[9] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott M. Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 1, 2

[10] Paolo Testolina, Francesco Barbato, Umberto Michieli, Marco Giordani, Pietro Zanuttigh, and Michele Zorzi. SELMA: Semantic large-scale multimodal acquisitions in variable weather, daytime and viewpoints. *arXiv preprint arXiv:2204.09788*, 2022. 1, 2

[11] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 2, 3

[12] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernández Domínguez. WildDash - Creating hazard-aware benchmarks. In *ECCV*, 2018. 1, 2

[13] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *CVPR*, 2022. 1, 2

[14] Jiaming Zhang, Kailun Yang, and Rainer Stiefelhagen. ISSAFE: Improving semantic segmentation in accidents by fusing event-based data. In *IROS*, 2021. 1

[15] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3D LiDAR semantic segmentation. In *ICCV*, 2021. 1