EVEN: An Event-Based Framework for Monocular Depth Estimation at Adverse Night Conditions

Peilun Shi[1], Jiachuan Peng[1], Jianing Qiu[1,2,†], Xinwei Ju[1], Frank Po Wen Lo[1], and Benny Lo[1]

***Abstract*— Accurate depth estimation under adverse night conditions has practical impact and applications, such as on autonomous driving and rescue robots. In this work, we studied monocular depth estimation at night time in which various adverse weather, light, and different road conditions exist, with data captured in both RGB and event modalities. Event camera can better capture intensity changes by virtue of its high dynamic range (HDR), which is particularly suitable to be applied at adverse night conditions in which the amount of light is limited in the scene. Although event data can retain visual perception that conventional RGB camera may fail to capture, the lack of texture and color information of event data hinders its applicability to accurately estimate depth alone. To tackle this problem, we propose an event-vision based framework that integrates low-light enhancement for the RGB source, and exploits the complementary merits of RGB and event data. A dataset that includes paired RGB and event streams, and ground truth depth maps has been constructed. Comprehensive experiments have been conducted, and the impact of different adverse weather combinations on the performance of framework has also been investigated. The results have shown that our proposed framework can better estimate monocular depth at adverse nights than six baselines.**

I. Introduction

Depth estimation with monocular cameras has been actively studied over the past decades [1], [2], [3], as it offers an efficient and economic way of obtaining depth. Compared to LiDAR, a monocular camera can be deployed pervasively, and due to its small scale, it can also be installed on an agent, e.g., an autonomous car, unobtrusively.

Albeit convenient and flexible, accurately estimating depth from a monocular camera is non-trivial, especially at night time, at which the visual perception of conventional RGB cameras degrades. The low dynamic range and sensitivity to motion blur of conventional cameras can lead to defective imaging at night, and the captured images/videos often exhibit underexposure due to low-lighting or back-lighting [4]. For an autonomous car, when it is driving at night accompanied by adverse weather (e.g., rain and fog), the dual occurrence of adverse light and weather can cause a challenge for its RGB-based vision system.

Recently, event camera has gained popularity in visual perception and robotics. Event camera is a bio-inspired vision sensor that works in a different way than conventional cameras [5], [6]. Rather than capturing intensity images at a fixed rate, event cameras measure intensity changes asynchronously in the form of an event stream. Event cameras have distinct advantages over conventional RGB cameras, including very high dynamic range (HDR), high temporal resolution, less motion blur, and low power consumption. These features of the event camera can complement its RGB counterpart, providing extra visibility and leading to an enhanced visual perception system.

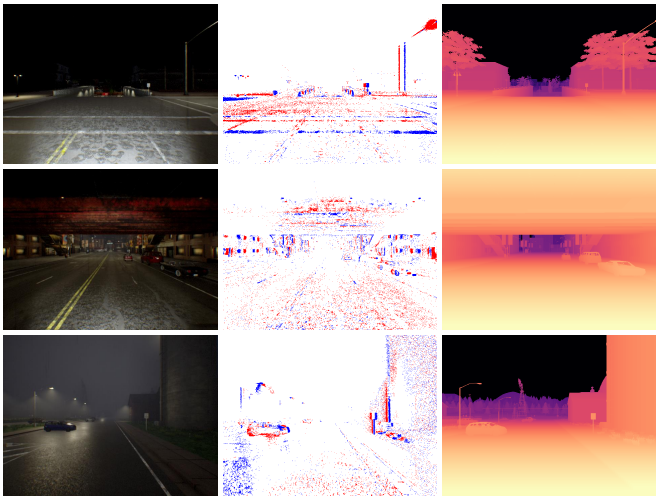


Fig. 1: Data samples from our MonoANC dataset. We show paired RGB and event images, and the ground truth depth map for each sample. The adverse night scenarios from top to bottom are: 1) driving in the heavy rain on a city road; 2) driving under a bridge at a foggy night; 3) driving at the countryside at a rainy and foggy night.

On the other hand, in depth estimation, texture and salient edges play more important roles than color as recognized by research in the computer vision community [7]. Texture can be well retained in RGB data whereas salient edges can be better captured by the event camera. Therefore, using both data modalities is a straightforward attempt to boost the overall depth estimation accuracy.

Although there are few studies [8], [9], [10] that have been proposed to jointly utilize RGB and event data for monocular depth estimation, they mainly focus on day time or normal weather conditions. Thus far, no research has been carried out on event-based monocular depth estimation under adverse night conditions, which is challenging as the RGB source does not contain as much effective visual information as it

†indicates the corresponding author
[1]The Hamlyn Centre, Imperial College London, London, U.K. {p.shi21, j.peng21, x.ju21, po.lo15, benny.lo}@imperial.ac.uk
[2]Department of Computing, Imperial College London, London, U.K. jianing.qiu17@imperial.ac.uk
The authors would like to thank Dr Lin Wang from Hong Kong University of Science and Technology, Guangzhou for his helpful discussion.

does at day time, and how to effectively fuse RGB data with event stream at night time has yet to be addressed.

Despite practical applications, such as more intelligent and lightweight night-time autonomous driving and rescue robots, there is currently also no dataset that contains paired RGB, event and ground truth depth data captured at adverse night conditions to validate and benchmark research in this direction. Hence, in this work, we made the following two contributions:

1) We propose the first adverse night-time driving dataset that contains paired RGB images, event streams, and ground truth depth maps. The adverse night conditions in our dataset are diverse in a variety of aspects including adverse weather such as rain and fog, and different scenes such as driving on dim countryside roads.
2) We propose a novel three-phase framework, which employs low-light enhancement and multi-modal fusion to tackle the problem of monocular depth estimation at adverse night conditions with event-based vision. The entire framework has been thoroughly evaluated, with the results showing that it outperforms six baselines.

II. Related Work

A. Monocular Depth Estimation with Multi-Modal Fusion

Monocular depth estimation can be achieved using RGB modality alone [1], [2], [3]. Recent advances in multiple data modalities have further improved the depth estimation accuracy. For instance, some research works proposed to use RGB and optical flow [11], [12], [13], [14], RGB combined with segmentation maps [15], [16], [17], or RGB with extra saliency features [18], [19] as the inputs, and use multi-modal fusion to enhance depth estimation.

LiDAR has been explored for enhancing monocular depth estimation recently. [20] and [21] proposed using late fusion methods to fuse depth data from LiDAR and monocular RGB inputs. Apart from pure visual signals, radar has also been used with RGB modality for monocular depth estimation [22], [23]. Recently, an attention-based method has been proposed for fusing radar signals with monocular RGB images [24].

讲了lidar与RGB的融合，可以尝试将lidar与事件相机融合

B. Event-Based Depth Estimation

第三篇 Daniel et al. [8] combined event-based data and monocular RGB frames with a recurrent asynchronous network for depth estimation, which is also the first work to fuse the event and monocular RGB frames. Zhou et al. [25] investigated the use of stereo event cameras for semi-dense depth estimation by maximizing a temporal consistency between the corresponding event streams. Another event vision-based method was proposed by Zhu et al. [9] which eliminates disparity for depth estimation. The method proposed by [10] shows the first learning-based stereo depth estimation for event cameras which is also the first one that produces dense results. [26] is an unsupervised framework that learns motion information only from event streams, achieving multi-task objectives including optical flow, egomotion and depth estimation. Cui et al. [27] proposed a dense depth estimation method based on the fusion of dense event stream and sparse point cloud.

下一篇看这个lidar与事件相机

Despite the efforts being made in event-based depth estimation, existing works are not engineered to specifically tackle monocular depth estimation at adverse night conditions, but instead mainly target at day time and normal weather conditions. In this work, we target monocular depth estimation at adverse night conditions. In order to improve the illumination in the field of view (FOV) and to take advantage of the HDR property of the event-based camera, we propose to combine low-light enhancement and multi-modal fusion of event and RGB data for better depth estimation. To the best of our knowledge, we are the first work that uses the event-based vision along with low-light image enhancement to estimate monocular depth at adverse night conditions.

是第一个使用基于事件的视觉和低光图像增强来估计不利夜间条件下的单眼深度的工作。

III. Method

Our framework decomposes monocular depth estimation at adverse night conditions into three phases as shown in Fig. 2. In phase one, the raw RGB image is first enlightened using low-light image enhancement; In phase two, the enhanced RGB image and the event image are fused to generate a fusion image; In phase three, depth estimation is carried out based on the fusion image. We denote our framework as **EVEN** as it is based on **EV**ent vision and low-light **EN**hancement. We elaborate our framework in the following.

A. Event Stream

将增强后的RGB图像与事件图像融合，生成融合图像

Asynchronous event streams reflect changes in light intensity. In order to efficiently make full use of the information from the event-based data. We convert event streams in the voxel grid format to image format. Specifically, spatial points (indexed by x and y positions in image coordinates with the value being the polarity p) are stacked along the time axis t using a fixed time period $\Delta t = 0.125$ s. This produces a compact event image.

B. Phase-1: Low-light Enhancement

The visual perception of conventional RGB cameras degrades at night due to the limited amount of light. To recover the necessary scene color and texture information captured by the RGB camera, we utilize EnlightenGAN [28] to enhance the raw night-time RGB image. EnlightenGAN is of an attention-based U-Net structure. The input RGB image is normalized by using the illumination channel as the attention map for the ultimate low-light enhancement.

C. Phase-2: Multi-modal Fusion

Event data can capture much more HDR and temporal details of night-time scenes, whereas RGB data can provide necessary texture and color information. As these two modalities complement each other, and in order to leverage the merits of both, a novel fusion network (refer to Fig. 3), which is built on top of selective kernel network [29], is designed to integrate event data with RGB modality.

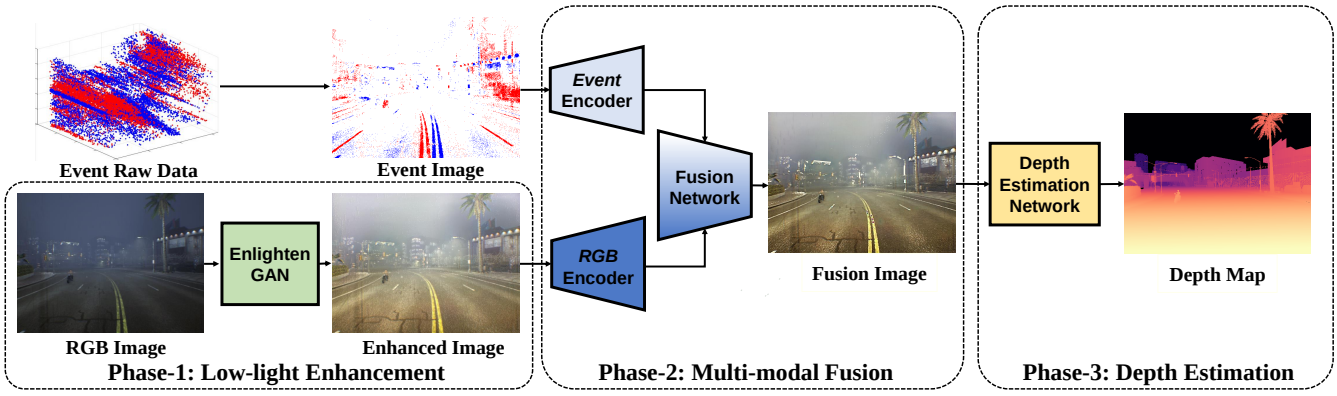


Fig. 2: An overview of the proposed framework for monocular depth estimation at adverse night conditions (e.g., at foggy night). Our framework, named EVEN, leverages a three-phase process to estimate depth: 1) phase-1: enlightening the low-light RGB image; 2) phase-2: fusing visual information from enhanced RGB and event images; 3) phase-3: estimating depth based on reconstructed fusion image.

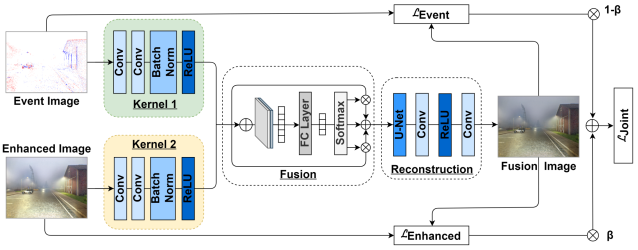


Fig. 3: The multi-modal fusion network of EVEN.

1) Fusion Network: given an event image $\mathbf{X}_{Event}$ and an enhanced RGB image $\mathbf{X}_{Enhanced}$, we use two convolutional kernels with different kernel sizes to transform the input images into feature maps. After transformation, two feature maps $\mathbf{F}_{Event}$ and $\mathbf{F}_{Enhanced}$ are obtained:

$$\mathbf{F}_{Event} = g(\mathbf{X}_{Event}), \mathbf{F}_{Event} \in \mathbb{R}^{H \times W \times C} \quad (1)$$

$$\mathbf{F}_{Enhanced} = h(\mathbf{X}_{Enhanced}), \mathbf{F}_{Enhanced} \in \mathbb{R}^{H \times W \times C} \quad (2)$$

where $g(\cdot)$ and $h(\cdot)$ are separate convolutional neural network layers that conduct transformation. For the event image, we use a kernel size of 5×5 as the information carried in event modality is relatively sparse. Therefore, a large kernel size is used. For the enhanced RGB image, we use a kernel size of 3×3. Following convolutional transformation, the feature maps of the two modalities are merged using an element-wise summation:

$$\mathbf{F}_{sum} = \mathbf{F}_{Event} + \mathbf{F}_{Enhanced}, \mathbf{F}_{sum} \in \mathbb{R}^{H \times W \times C} \quad (3)$$

We then apply global average pooling to conduct dimension reduction (along the H and W dimensions) for the merged feature map $\mathbf{F}_{sum}$, which produces a vector $\mathbf{V} \in \mathbb{R}^{1 \times C}$. Similar to [29], we then use a simple fully connected layer $f(\cdot)$ to create a compact vector $\mathbf{k}$ on the basis of $\mathbf{V}$:

$$\mathbf{k} = f(\mathbf{V}), \mathbf{k} \in \mathbb{R}^{d \times 1} \quad (4)$$

$\mathbf{k}$ is then used to guide adaptive fusion of the two modalities. Specifically, we create soft attention across channel C. For c-th element along the channel C, the soft attention for fusing event and enhanced RGB feature maps can be formulated as follows:

$$a_c = \frac{e^{\mathbf{A}_c \mathbf{k}}}{e^{\mathbf{A}_c \mathbf{k}} + e^{\mathbf{B}_c \mathbf{k}}}, b_c = \frac{e^{\mathbf{B}_c \mathbf{k}}}{e^{\mathbf{A}_c \mathbf{k}} + e^{\mathbf{B}_c \mathbf{k}}} \quad (5)$$

$$\mathbf{F}_{fused_c} = a_c \cdot \mathbf{F}_{Event_c} + b_c \cdot \mathbf{F}_{Enhanced_c}, a_c + b_c = 1 \quad (6)$$

where $\mathbf{A}_c \in \mathbb{R}^{1 \times d}$ and $\mathbf{B}_c \in \mathbb{R}^{1 \times d}$ are learnable vectors.

The fused feature map $\mathbf{F}_{fused}$ is then fed into an U-Net [30] followed by a group of convolution and ReLU operations to 1) further fuse features of the event and RGB modalities, and 2) reconstruct a fusion image $\mathbf{Y}$ of the same resolution to the input event and enhanced RGB images:

$$\mathbf{Y} = \text{Conv}(\text{ReLU}(\text{Conv}(\text{U-Net}(\mathbf{F}_{fused})))) \quad (7)$$

The resulting fusion image, which has HDR property and better edge salience, also suppresses areas of overexposure caused by low-light enhancement as shown in Fig. 3.

2) Fusion Loss: In order to allow the entire fusion network to effectively merge visual information from the two modalities, a joint loss $\mathcal{L}_{joint}$ is designed as shown in Equation 8. We use the reconstruction loss between the fusion image and the enhanced RGB image as the primary loss (i.e., $\mathcal{L}_{Enhanced}$), and that between the fusion image and the event image as the auxiliary loss (i.e., $\mathcal{L}_{Event}$). Both reconstruction losses are implemented as an $\mathcal{L}_2$ loss that measures the mean squared error between the fusion image and the respective event or enhanced RGB image. During training, the fusion network is trained to decrease $\mathcal{L}_{joint}$.

$$\mathcal{L}_{joint} = \beta \times \mathcal{L}_{Enhanced} + (1 - \beta) \times \mathcal{L}_{Event} \quad (8)$$

D. Phase-3: Depth Estimation

The fusion image, which contains visual information from both event and RGB modalities, is then used as the source for depth estimation. We separately adopt two state-of-the-art depth estimation networks, i.e., Depthformer [31] and SimIPU [32] in our EVEN framework to carry out the depth estimation with the fusion image as their input.

IV. Dataset

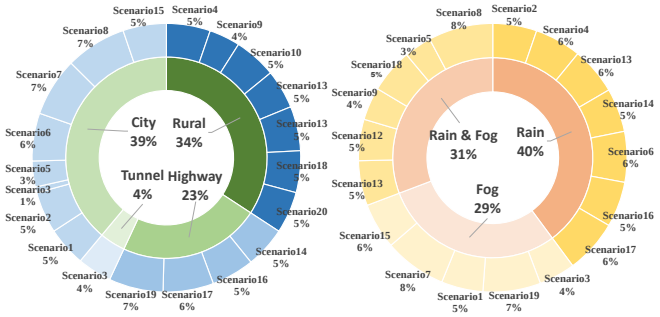
To the best of our knowledge, there is currently no dataset that is proposed for monocular depth estimation at adverse night conditions, containing paired RGB, event and depth images. In order to validate the effectiveness of our proposed framework, and advance future research in this direction, we construct the first adverse night-time driving dataset that includes the aforementioned data modalities and the ground truth depth maps. The dataset was constructed using CARLA [33], a popular simulator in the field of autonomous driving, and the event camera plugin [34].

A. Data Collection and Statistics

We collect the data through a sensor suite that contains an RGB camera, an event-based camera, and a depth camera. The positive and negative thresholds for triggering an event of the event camera was set to 0.4. All the sensors were set with a FOV of 90 degrees, a resolution of 640×480 pixels, and a data generation rate of 8 Hz. All sensors had an egocentric view mounted on a vehicle while it was driving around. The start and end points of the vehicle were manually selected and the routes of the vehicle were accurately planned. The speed and following distance of the vehicles as well as the lights at night were based on road traffic guidelines to achieve the maximum realism of night-time driving. The driving scenes are diverse, including typical city and rural roads, as well as tunnel and highway. The statistics of the driving scenarios is shown in Fig. 4a. We also apply adverse weather conditions such as rain, fog, and the combination of rain and fog to the scene, and Fig. 4b shows the distribution of the adverse weather in our dataset. The dataset contains 11,191 samples. Each sample has paired RGB, event and ground truth depth images. We split the entire dataset into 70% for training, 15% for validation and the rest 15% for testing. We name our dataset as **MonoANC** (**Mono**cular depth estimation at **A**dverse **N**ight **C**onditions).

V. Experiment

In this section, we first describe the implementation details of our framework - EVEN, and then the evaluation metrics, followed by the baseline methods that are used to compare against our framework. We then show overall results of all methods on MonoANC, and present the results of cross validation of the performance of EVEN on different adverse weather combinations at the end.



(a) (b)

Fig. 4: Distribution of different night-time driving environments (a) and different adverse weather conditions (b).

A. Implementation Details

We implement our EVEN framework using PyTorch. The learning rate for training the multi-modal fusion network was 1e-3. AdamW [35] was used as the optimizer. Weight decay was set to 1e-3. Step size of scheduler was 5 during the training of the fusion network and we trained it for 100 epochs. We set β to 0.8 in Equation 8. After the fusion network was properly trained, we pre-generated the fusion images, and trained depth estimation network (i.e., Depthformer and SimIPU) using their default settings [36].

B. Evaluation Metrics

We use standard evaluation protocols following [37] to evaluate our framework and all baseline methods. Specifically, we measure the mean absolute relative error (Abs. Rel.), mean squared relative error (Sq. Rel.), root mean squared error (RMSE), and mean log10 error (Log10). Apart from the error metrics, we also adopted three different thresholds as the accuracy metrics which are the common practice in the literature, i.e., $\alpha = 1.25^i, i = 1, 2, 3$.

C. Baseline Methods

We implement six baselines to compare and examine the effectiveness of our framework on boosting depth estimation, i.e., the use of low-light enhancement and fusion with event and RGB modalities. As mentioned early, salient edges and texture are core features for depth estimation. We therefore adopted the Sobel operator [38], which is an edge detector, to process the RGB modality, and using the resulting image as the alternative to the event image in our framework to justify the use of event data, which is also able to retain salient edge information.

1) RGB: the raw RGB image is fed directly into the depth estimation network as the only input for depth estimation.
2) Event: the event image is fed directly into the depth estimation network as the only input.
3) RGB + Sobel: the paired raw RGB and Sobel operator processed images are used as the inputs to the phase-2 of EVEN, followed by depth estimation of phase-3.

TABLE I: Results on MonoANC Dataset When the Depth Estimation Network in EVEN is Instantiated as Depthformer and SimIPU Respectively

Input Sequence	Depthformer							SimIPU						
	Error Metric ↓				Accuracy Metric ↑			Error Metric ↓				Accuracy Metric ↑		
	Abs. Rel.	Sq. Rel.	RMSE	Log10	$\alpha 1$	$\alpha 2$	$\alpha 3$	Abs. Rel.	Sq. Rel.	RMSE	Log10	$\alpha 1$	$\alpha 2$	$\alpha 3$
RGB	0.192	0.310	4.973	0.069	0.810	0.911	0.985	0.293	0.370	5.177	0.079	0.710	0.921	0.972
Event	0.452	0.220	7.775	0.172	0.390	0.622	0.795	0.594	1.240	9.180	0.116	0.552	0.828	0.932
RGB + Sobel	0.180	0.340	5.304	0.064	0.808	0.908	0.956	0.266	0.310	4.947	0.067	0.773	0.930	0.976
RGB + Event	0.179	0.340	5.992	0.067	0.795	0.920	0.956	0.229	0.280	5.151	0.057	0.837	0.953	0.984
$\text{RGB}_{Enhanced}$	0.181	0.390	5.737	0.074	0.765	0.924	0.971	0.263	0.300	4.998	0.058	0.824	0.948	0.984
$\text{RGB}_{Enhanced}$ + Sobel	0.139	**0.280**	5.023	0.063	0.806	0.970	0.988	0.216	**0.240**	**4.080**	0.063	0.846	0.954	0.986
EVEN (Ours)	**0.112**	**0.280**	**4.335**	**0.049**	**0.903**	**0.976**	**0.993**	**0.125**	0.280	4.845	**0.049**	**0.857**	**0.959**	**0.988**

TABLE II: Cross Validation Results of EVEN on Different Adverse Weather Conditions

Input Sequence		Depthformer							SimIPU						
		Error Metric ↓				Accuracy Metric ↑			Error Metric ↓				Accuracy Metric ↑		
Train Set	Test Set	Abs. Rel.	Sq. Rel.	RMSE	Log10	$\alpha 1$	$\alpha 2$	$\alpha 3$	Abs. Rel.	Sq. Rel.	RMSE	Log10	$\alpha 1$	$\alpha 2$	$\alpha 3$
rain and fog at the same time	rain only and fog only	0.325	1.987	8.475	0.187	0.471	0.645	0.797	0.330	1.865	8.710	0.187	0.420	0.655	0.786
rain only and fog only	rain and fog at the same time	**0.267**	**0.315**	**4.934**	**0.031**	**0.646**	**0.833**	**0.937**	**0.260**	**0.307**	**4.933**	**0.031**	**0.680**	**0.844**	**0.939**

4) RGB + Event: the paired raw RGB and event images are used as the inputs to the phase-2 of EVEN, followed by depth estimation of phase-3.
5) $\text{RGB}_{Enhanced}$: the enhanced RGB image after phase-1 is fed directly into the depth estimation network as the only input for depth estimation.
6) $\text{RGB}_{Enhanced}$ + Sobel: the paired enhanced RGB image after phase-1 and Sobel operator processed image are used as the inputs to the phase-2 of EVEN, followed by depth estimation of phase-3.

D. Overall Results

As we instantiate the depth estimation network separately as either a Depthformer or a SimIPU, we run six baselines accordingly based on the instantiated depth estimation network. Table I summarizes the overall results. Our complete EVEN framework outperforms the baseline methods, and its performance improvement is consistent across Depthformer and SimIPU. The absolute relative error (Abs. Rel.) is reduced by 41.7% and 57.3% respectively compared to a single RGB input to the Depthformer and SimIPU. An 11.5% relative improvement on $\alpha 1$ accuracy metric can also be observed for EVEN using Depthformer, and a 20.7% increase for EVEN using SimIPU, compared to a single RGB image as the input to these two depth estimation networks.

Fig. 5 shows four qualitative results of depth estimation on MonoANC. It can be observed that the depth maps estimated by EVEN have a noticeable improvement in detail at the edges as well as the objects in the far distance compared to those of baselines. As indicated by the red boxes in Fig. 5, our complete EVEN framework can produce depth maps without much artifacts, and are closer to the ground truth. These visually prove that the fusion of the edge information and HDR features of event data in EVEN is effective. When we replace the event image with Sobel operator processed image, i.e., indicated by $\text{RGB}_{Enhanced}$ + Sobel, the quality of the estimated depth map slightly degrades, but is still better than those of the rest baseline methods.

E. Cross Validation on Adverse Weather

We further split MonoANC based on different weather conditions. Specifically, there are three adverse weather conditions as shown in Fig. 4(b): 1) rain only; 2) fog only; 3) rain and fog occur together. We split the dataset into two sets. One set contains samples of rain only and fog only, and the other set contains samples of simultaneous occurrence of rain and fog in the scene. A two-fold cross-validation is then conducted to evaluate the performance of EVEN. Table II shows the results. When the framework has seen each individual weather condition during the training, it can well estimate the depth of the scene with mixed adverse weather conditions, i.e., rain and fog occurring at the same time in the scene. Conversely, it becomes difficult for the framework to estimate depth for the scenes with only a single adverse weather condition if the training data is scenes of mixed adverse weather. Hence, cost function of decomposing adverse weather combinations is worth investigating for better depth estimation in future work.

VI. Conclusion

In this paper, we have proposed a framework that integrates low-light enhancement and fuses RGB and event modailies for effective monocular depth estimation under adverse night conditions. A synthetic night-time driving dataset that contains paired RGB, event and depth images has also been constructed, which includes scenes that encountering adverse weather, light and road conditions. The experiment results have shown that our proposed framework is able to achieve satisfactory depth estimation results in various adverse night scenarios.

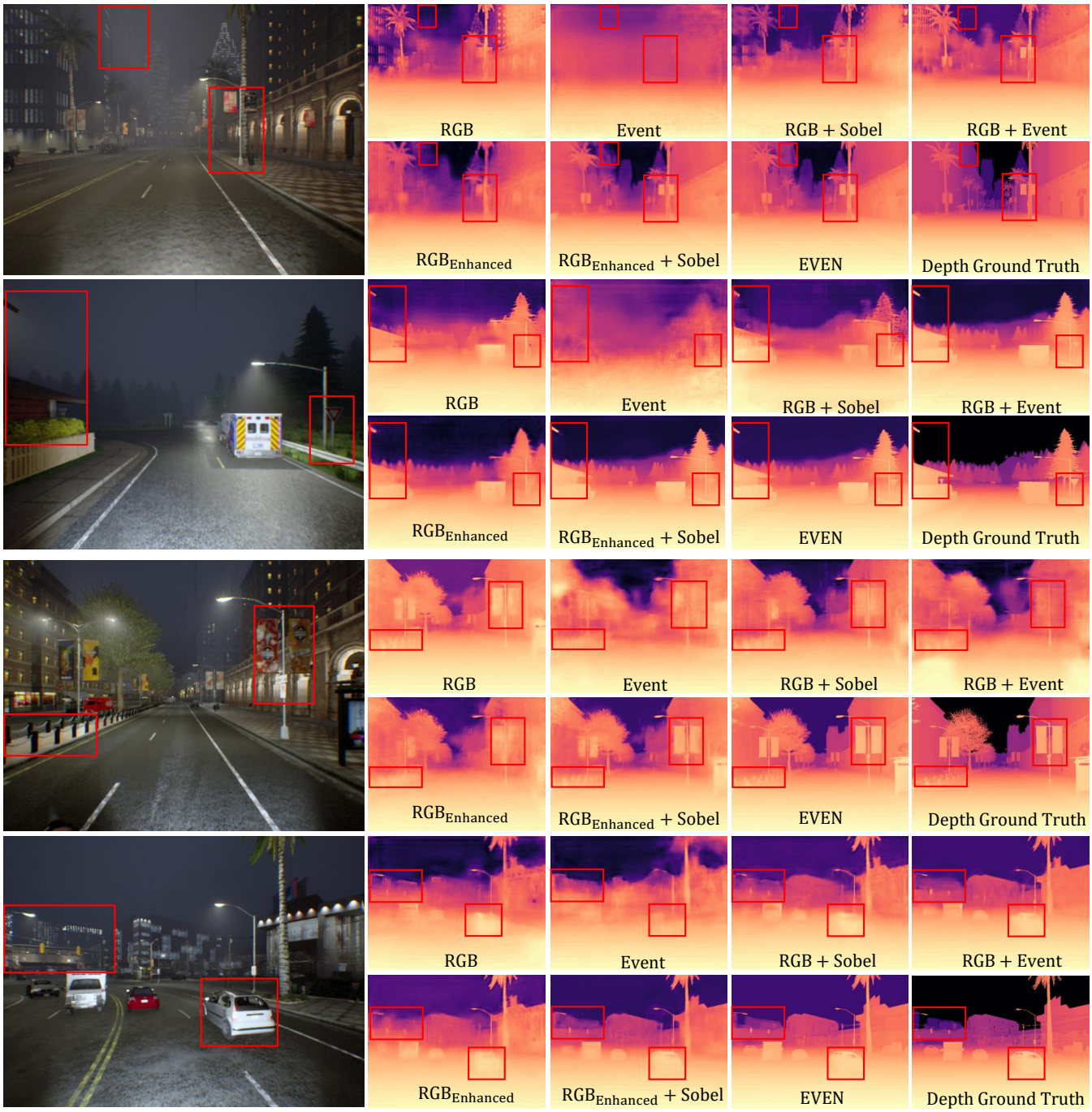


Fig. 5: Qualitative results of depth estimation on MonoANC dataset. Top two examples are the results when Depthformer is adopted as the depth estimation network, and the bottom two examples are the results when SimIPU is adopted as the depth estimation network. Areas indicated by the red boxes show that our EVEN framework can better estimate monocular depth than other baseline methods.

REFERENCES

[1] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[2] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.

[3] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8001–8008.

[4] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6849–6857.

[5] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis *et al.*, "Event-based vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 154–180, 2020.

[6] P. Lichtsteiner and C. Posch, "C. and t. delbruck,"an 128× 128 120 db 15 μs latency temporal contrast vision sensor," *IEEE Journal of Solid State Circuits*, vol. 43, pp. 566–576, 2008.

[7] J. Hu, Y. Zhang, and T. Okatani, "Visualization of convolutional neural networks for monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3869–3878.

[8] D. Gehrig, M. Rüegg, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2822–2829, 2021.

[9] A. Z. Zhu, Y. Chen, and K. Daniilidis, "Realtime time synchronized event-based stereo," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[10] S. Tulyakov, F. Fleuret, M. Kiefel, P. Gehler, and M. Hirsch, "Learning an event sequence embedding for dense event-based deep stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1527–1537.

[11] R. Ranftl, V. Vineet, Q. Chen, and V. Koltun, "Dense monocular depth estimation in complex dynamic scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4058–4066.

[12] K. Tateno, F. Tombari, I. Laina, and N. Navab, "Cnn-slam: Real-time dense monocular slam with learned depth prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6243–6252.

[13] J. Chen, X. Yang, Q. Jia, and C. Liao, "Denao: Monocular depth estimation network with auxiliary optical flow," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 8, pp. 2598–2610, 2020.

[14] T. Shimada, H. Nishikawa, X. Kong, and H. Tomiyama, "Pix2pix-based monocular depth estimation for drones with optical flow on airsim," *Sensors*, vol. 22, no. 6, p. 2097, 2022.

[15] P. Zama Ramirez, M. Poggi, F. Tosi, S. Mattoccia, and L. Di Stefano, "Geometry meets semantics for semi-supervised monocular depth estimation," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 298–313.

[16] S. Zhu, G. Brazil, and X. Liu, "The edge of depth: Explicit constraints between segmentation and depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 116–13 125.

[17] L. He, J. Lu, G. Wang, S. Song, and J. Zhou, "Sosd-net: Joint semantic object segmentation and depth estimation from monocular images," *Neurocomputing*, vol. 440, pp. 251–263, 2021.

[18] Y.-f. Zhang, J. Zheng, W. Jia, W. Huang, L. Li, N. Liu, F. Li, and X. He, "Deep rgb-d saliency detection without depth," *IEEE Transactions on Multimedia*, vol. 24, pp. 755–767, 2021.

[19] S. Abdulwahab, H. A. Rashwan, M. A. Garcia, A. Masoumian, and D. Puig, "Monocular depth map estimation based on a multi-scale deep architecture and curvilinear saliency feature boosting," *Neural Computing and Applications*, pp. 1–18, 2022.

[20] M. Jaritz, R. De Charette, E. Wirbel, X. Perrotton, and F. Nashashibi, "Sparse and dense data with cnns: Depth completion and semantic segmentation," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 52–60.

[21] C. Fu, C. Mertz, and J. M. Dolan, "Lidar and monocular camera fusion: On-road depth completion for autonomous driving," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 273–278.

[22] J.-T. Lin, D. Dai, and L. Van Gool, "Depth estimation from monocular images and sparse radar data," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 233–10 240.

[23] C.-C. Lo and P. Vandewalle, "Depth estimation from monocular images and sparse radar using deep ordinal regression network," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3343–3347.

[24] J. Long, J. Huang, and S. Wang, "Radar fusion monocular depth estimation based on dual attention," in *International Conference on Adaptive and Intelligent Systems*. Springer, 2022, pp. 166–179.

[25] Y. Zhou, G. Gallego, H. Rebecq, L. Kneip, H. Li, and D. Scaramuzza, "Semi-dense 3d reconstruction with a stereo event camera," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 235–251.

[26] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 989–997.

[27] M. Cui, Y. Zhu, Y. Liu, Y. Liu, G. Chen, and K. Huang, "Dense depth-map estimation based on fusion of event camera and sparse lidar," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[28] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE Transactions on Image Processing*, vol. 30, pp. 2340–2349, 2021.

[29] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[31] Z. Li, Z. Chen, X. Liu, and J. Jiang, "Depthformer: Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation," *arXiv preprint arXiv:2203.14211*, 2022.

[32] Z. Li, Z. Chen, A. Li, L. Fang, Q. Jiang, X. Liu, J. Jiang, B. Zhou, and H. Zhao, "Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations," *arXiv preprint arXiv:2112.04680*, 2021.

[33] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[34] J. Hidalgo-Carrió, D. Gehrig, and D. Scaramuzza, "Learning monocular dense depth from events," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 534–542.

[35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[36] Z. Li, "Monocular depth estimation toolbox," https://github.com/zhyever/Monocular-Depth-Estimation-Toolbox, 2022.

[37] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[38] N. Kanopoulos, N. Vasanthavada, and R. L. Baker, "Design of an image edge detection filter using the sobel operator," *IEEE Journal of solid-state circuits*, vol. 23, no. 2, pp. 358–367, 1988.