2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Improved Event-Based Dense Depth Estimation via Optical Flow Compensation

Dianxi Shi[1,2], Luoxi Jing[1], Ruihao Li[2], Zhe Liu[3], Lin Wang[3], Huachi Xu[2], Yi Zhang[2]

***Abstract*— Event cameras have the potential to overcome the limitations of classical computer vision in real-world applications. Depth estimation is a crucial step for high-level robotics tasks and has attracted much attention from the community. In this paper, we propose an event-based dense depth estimation architecture, Mixed-EF2DNet, which firstly predicts inter-grid optical flow to compensate for lost temporal information, and then estimates multiple contextual depth maps that are fused to generate a robust depth estimation map. To supervise the network training, we further design a smoothing loss function used to smooth local depth estimates and facilitate estimating reasonable depth for pixels without events. In addition, we introduce SE-resblocks in the depth network to enhance the network representation by selecting feature channels. Experimental evaluations on both real-world and synthetic datasets show that our method performs better in terms of accuracy when compared to state-of-the-art algorithms, especially in scene detail estimation. Besides, our method demonstrates excellent generalization in cross-dataset tasks.**

I. Introduction

Bio-inspired dynamic vision sensors called event cameras, such as Dynamic Vision Sensor (DVS) [1], are novel event-driven devices that only report the pixel where illumination intensity has changed beyond a set threshold. Unlike conventional sensors that capture a global intensity image, as shown in Fig. 1, event cameras output asynchronous event streams, which consist of positive and negative events that indicate the increase or decrease in illumination, respectively. Each event contains the timestamp it occurs, the pixel-location, and polarity information of brightness changes. With the advantages of high temporal resolution, high dynamic range, low power consumption, and no motion blur, event cameras offer a potential choice for the scenarios where conventional cameras are challenged [2], [3].

Depth estimation aims to provide reasonable explanations of the observed scene, which is fundamental and of great importance to a variety of high-level computer vision applications such as autonomous driving and robotic grasping [4]–[6]. Different from classical computer vision tasks, event-based dense depth estimation takes event stream instead of intensity images to estimate the depth for each pixel. It is well suited for scenarios with poor lighting conditions, battery limitations, or where high-speed responses are required,

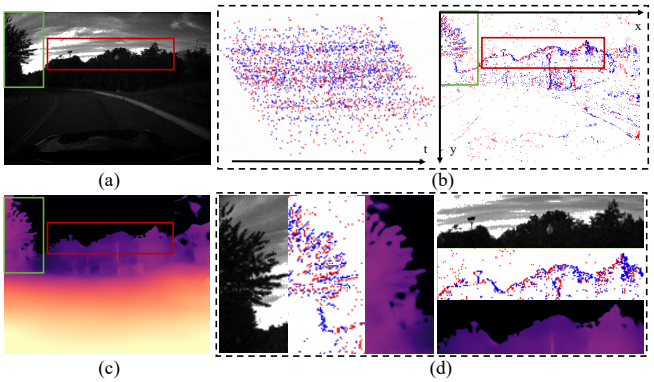


Fig. 1. Qualitative results of our proposed framework. (a) Intensity image. (b) Event stream. Red and blue dots indicate positive and negative events, respectively. (c) Dense depth map predicted by our method. The brighter color indicates less depth. (d) Details of our prediction.

such as obstacle avoidance [7]. However, the event stream is sparse and asynchronous by nature, meaning there are missing values for some pixels over a while. This makes the real-time pixel-wise dense depth estimation challenging.

In this paper, we propose a dense depth estimation method based on a monocular event stream that accurately estimates the log depth of the scene, especially the details of objects. It predicts global depth robustly even in challenging environments and shows satisfied generalization on across-data tasks. Our method first constructs a voxel grids list from events as the input of network to increase the available scene information for single-step training. For an input list, multiple flow compensation features are constructed to fuse optical flow with voxel grids, which compensates for the lost temporal correlation of the event stream during voxelization and enhances the object contours to help with scene understanding. Finally, our network estimates multiple contextual depth maps that are fused to generate a final estimation map corresponding to the interest region of the input list. The qualitative results of our proposed method are shown in Fig. 1 (c) and (d). Our contributions can be summarized as follow:

- We propose an event-based monocular dense depth estimation framework called Mixed-EF2DNet. The proposed Mixed-EF2DNet compensates inputs with estimated optical flow and predicts multiple contextual depth maps that are fused to improve the accuracy and robustness of depth estimation.
- We design a local smoothness loss function applied on predicted depth maps to facilitate reasonably estimated depths for pixels where no events occur. Besides, SE blocks are introduced into residual modules to recalibrate feature maps and improve network representation.

*This work was supported by the National Natural Science Foundation of China (Grant No. 91948303, No. 61903377).

[1]School of Computer Science, Peking University, Beijing, 100871, China.
[2]Artificial Intelligence Research Center (AIRC), Defense Innovation Institute, Beijing 100166, China.
[3]College of Computer, National University of Defense Technology, Changsha, 410073, China.
Correspondence: jxnzdl@126.com

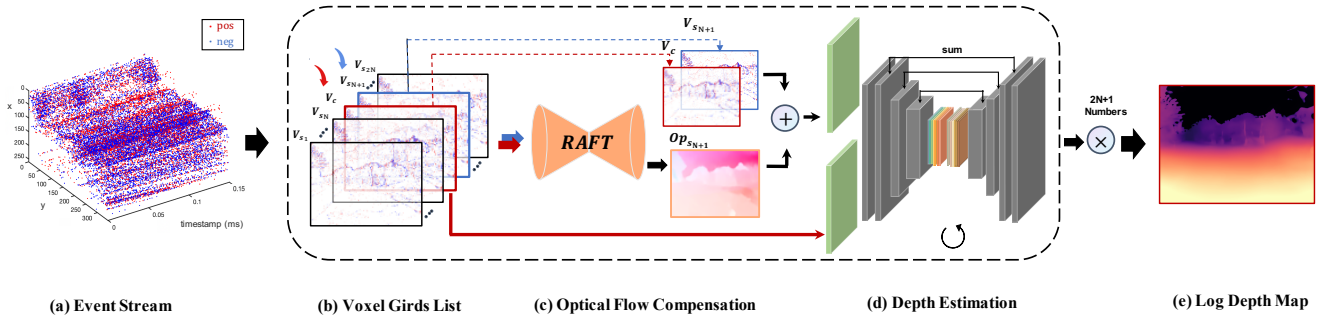


Fig. 2. Overview of our proposed system. (a) Input event stream. (b) Construct input voxel grid list of length 2N+1, containing one central voxel grid V_c (marked by a red box) and 2N side voxel grids $\{V_{s_i}\}_{i=1}^{2N}$. (c) Build flow compensation features for each side voxel grid. The side voxel grid $V_{s_{N+1}}$ (marked by a blue box), for example, with the central voxel grid V_c are fed into the optical flow network to get a flow map $Op_{s_{N+1}}$ (marked by an orange box). The flow map $Op_{s_{N+1}}$ concatenated with the input two grids V_c and $V_{s_{N+1}}$ forming a flow compensation feature. (d) Depth estimation. Feed the 2N flow compensation features and the central voxel grid separately into depth estimation to get 2N+1 contextual depth estimations, which are fused to get the final depth map. (e) Get the log depth estimation corresponding to the grid of interest V_c

- We implement the proposed system and perform corresponding experiments on both real-world and synthetic datasets, with comparisons against state-of-the-art methods. The results show that our framework outperforms baseline methods in both days and challenging night environments with outstanding generalization.

II. Related Work

Classical depth estimation has been researched and made excellent results in the last decade [8]–[10]. However, these standard computer vision techniques cannot be directly applied to unstructured three-dimensional event stream that is fundamentally different from standard images. Recently, event-based depth estimation methods have attracted considerable attention from the community and can be divided into model-based methods and learning-based methods.

Model-based methods underwent exploration from sparse to dense depth estimation [11]–[14]. These methods are typically demonstrated in scenes with strong assumptions resulting in poor generalization. Learning-based methods significantly improve the performance of depth estimation by large margins. Zhu et al. [15] proposed an unsupervised network that jointly estimates depth, optical flow, and ego-motion only from the event stream. However, the predicted depth maps are still semi-dense. Tulyakov et al. [16] proposed an event sequence embedding module on which a deep neural network was designed to perform dense depth estimation from the stereo event stream. Besides, Gehrig et al. [17] used events and intensity images to complement each other and further estimated dense depth maps. They proposed a recurrent asynchronous multimodal network to fuse data taken by various types of measurement sensors. Stereo setting or multiple type sensors systems with complex structures require more resources and computation than the monocular setting, which can be a disadvantage for tasks with energy limitations. The most related to our method is [18], where a recurrent encoder-decoder style framework is used to perform dense depth estimation from monocular event stream by leveraging the temporal consistency. Our method solves the same task by a supervised joint network that further introduces optical flow compensating temporal relationship lost by event pre-processing than [18]. Unlike [15], where optical flow is calculated from depth predictions and then used as a loss function to supervise the depth network training, our method uses optical flow as a part of the input to depth estimation for providing temporal information and scene features.

III. Method

In this section, we describe our approach for estimating dense depth maps from a given monocular asynchronous event stream. We start by describing the processing of event stream into image-like representation. Then we give an overview of our framework and introduce detailed compositions of the proposed Mixed-EF2DNet. Finally, we provide the designed loss functions driving the network training.

A. Event Representation

For event cameras, pixels remain independent of each other and respond to changes of intensity asynchronously. An event $e = (x, y, t, p)$ is triggered at pixel location (x, y) and timestamp t if the change in log intensity since the last event is greater than the threshold. Polarity $p = \pm 1$ denotes the direction of the intensity change.

Given high temporal resolution, event cameras allow a large number of events to occur and form a sparse event stream in a short period. We encode events in the voxel grid [15] so that they can be fed to the neural network. We divided input event stream E into event windows $\{\varepsilon\}$ by a fixed duration ΔT. For each event window $\varepsilon_k = \{e_i\}_{i\in[1,\mathrm{M}]}$ with M events, the range of timestamps is divided into B bins, and timestamps are scaled to the range $[0, B-1]$. We generate event volume V_k corresponding to event window ε_k by linearly weighted accumulation. The event volume V_k with dimensions $B \times H \times W$ is defined as below:

$$V_k(x,y,t) = \sum_{e_i} p_i \delta(x - x_i, y - y_i) max(0, 1 - |t - t_i^*|) \quad (1)$$

where $t_i^* = \frac{(B-1)}{\Delta T}(t_i - t_0)$ denotes the normalized timestamp.

B. Network Architecture

Our Mixed-EF2DNet takes a voxel grid list generated from the event stream and outputs a logarithmic depth map for the grid of interest. It consists of an optical flow network and depth network. An overview of our approach is given in Fig.

2. The optical flow network estimates flow maps between voxel grids. Then the flow maps are fused with inputs by constructing flow compensation features. The depth network takes flow compensation features or the grid of interest to estimate multiple contextual depth maps that are fused to get a robust final depth map. We will give the detailed description of our system in the following.

A single voxel grid generated from the event stream is sparse and contains part of the scene due to the principle of event cameras, which is challenging to estimate dense depth maps. To mitigate this problem, we construct a voxel grid list as input to increase the scene information used by Mixed-EF2DNet. The input voxel grid list $\{V_m\}_{m=1}^{2N+1}$ is continuous and non-overlapping in temporal dimension. It consists of one central voxel grid V_c, which is the interest, and 2N side voxel grids V_s that lie symmetrically before and after V_c. A schematic of the input voxel grids list is shown in Fig. 2 (b).

While the image-like representation such as voxel grid could be directly taken by neural networks, it sacrifices the high temporal resolution of original event stream. Thus, we introduce an optical flow network to estimate the flow map representing per-pixel motion between voxel grids for temporal compensation. Besides, the predicted flow maps with object contours could help depth estimation to understand the scene. We construct flow compensation features to fuse optical flow with inputs grids. Specifically, our optical flow network estimates flow maps $\{Op_{s_i}\}_{i=1}^{2N}$ from the central voxel grid V_c to each side voxel grid V_{s_i}. Every predicted flow map Op_s is concatenated with the two input voxel grids V_c and V_{s_i} to form a flow compensation feature CF_{s_i}. This process is illustrated in Fig. 2(c). We obtain a total of 2N flow compensation features for an input voxel grid list finally.

Then, the depth network of Mixed-EF2DNet takes flow compensation features in turn and outputs 2N contextual depth maps $\{D_{s_i}\}_{i=1}^{2N}$. In addition, the central voxel grid is also fed into the depth network and gets a contextual depth map D_c. Taking the central voxel grid as one input data for the depth network increases the weight of interest and suppresses the errors introduced by optical flow. This process is shown in Fig. 2(d).

Finally, Mixed-EF2DNet weights the 2N+1 contextual depth maps and sums them up to get a final depth map that corresponds to central voxel grid V_c, i.e, the grid of interest. Fusing multiple depth maps reduces the effect of extreme predicted values and enhances the robustness of the final depth map. The final depth map D_{final} defined as below:

$$D_{final} = \lambda_c D_c + \sum_{i=1}^{2N} \lambda_s D_{s_i} \tag{2}$$

where λ_c and λ_s present weights of contextual depth maps.

Optical Flow Network. To quickly verify the effectiveness of introducing optical flow for our task, we use a mature image-based optical flow network RAFT [19] to estimate the optical flow between voxel grids. A pre-trained model without fine-tuning was applied directly. The process is illustrated in Fig. 3. The central voxel grid V_c and a side voxel grid V_{s_i} are passed into RAFT to output the optical flow Op_{s_i} from the central voxel grid to the side one.

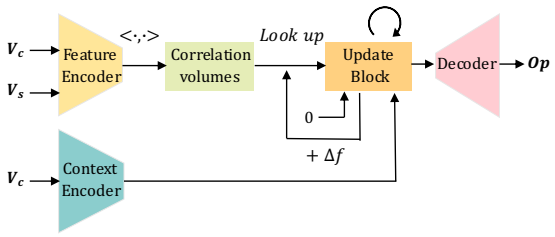


Fig. 3. The architecture of optical flow network. We use a mature image-based optical flow network RAFT here, including three stages: feature extraction, visual similarity computation, and iterative updates.

Depth Network. The depth network of Mixed-EF2DNet is a recurrent network based on UNet architecture [20]. Our depth network designs an input layer to adapt different input data. Furthermore, it introduces SE-resblock to recalibrate channel-wise feature responses, and improve the network representation by focusing on important information. The detailed architecture of the depth network is shown in Fig. 4. The input layer contains two convolutions with different input channels to process flow compensation features and the central voxel grid, respectively. Each encoder layer consists of a downsampling convolution followed by a ConvLSTM [21] that maintains and updates a state every iteration. In each SE-resblock, the SE block [22] performs a squeeze and excitation operation on the feature map U output by two convolutions and gets a collection of per-channel modulation weights W. Scale the product of feature map U and weights W and add the input of SE-resblock to the result to get an output feature map. Decoder layers upsample the feature map to full resolution by bilinear interpolation followed by a convolution. We use summation for skip connection between symmetric layers of encoders and decoders to fuse features of different levels.

C. Loss Function

We proposed an objective function to train Mixed-EF2DNet in a supervised fashion. Following previous works, we use the scale-invariant loss and multi-scale scale-invariant gradient matching loss as terms of the objective function. As depth is spatially continuous, neighboring pixels should have similar depths. Based on this insight, we designed a local smoothness loss term applied on the predicted depth map to smooth local depths while facilitating the pixel without an event to produce more reasonable depth estimates. The objective function is defined as below:

$$L_{total} = \sum_{k=0}^{L-1} \lambda_1 L_{k,si} + \lambda_2 L_{k,grad} + \lambda_3 L_{smooth} \tag{3}$$

where L denotes the number of unrolling steps for training the network, and λ_1, λ_2 and λ_3 are hyper-parameters. We use the normalized log depth maps for loss calculation to learn larger depth variations. Define log depth difference $R_k = \widehat{D_k} - D_k$, which denotes the difference between log ground truth $\widehat{D_k}$ and log depth predictions D_k. The scale-invariant loss is defined as below:

$$L_{k,si} = \frac{1}{n}\sum_u (R_k(u))^2 - \frac{1}{n^2}(\sum_u R_k(u))^2 \tag{4}$$

where $u = \{e_j\}_{j\in[1,2,..,n]}$ whose number is n denotes the valid pixels of ground truth. The multi-scale scale-invariant

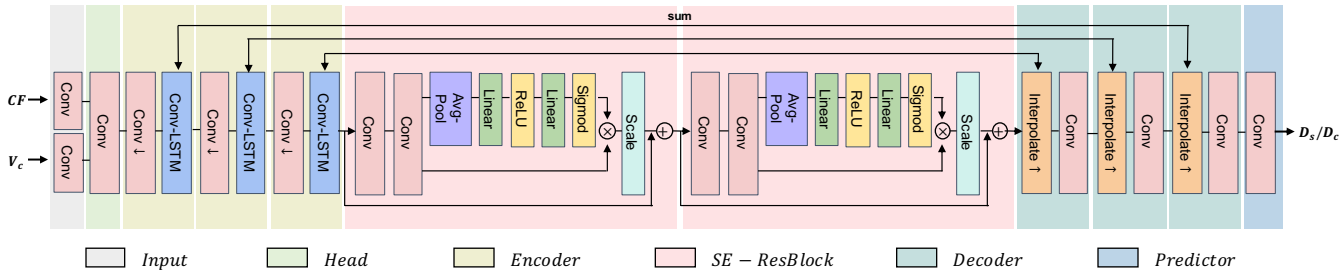


Fig. 4. The architecture of depth network. The depth network based on UNet architecture takes a flow compensation feature CF_{s_i} or the central voxel grid V_c to output a contextual depth map. Every SE-resblock contains two convolutions followed by a Squeeze-and-Excitation(SE) block. The *squeeze* operator performs a global average pooling on the feature map U output by the two convolution layers. Then, the *excitation* operator outputs a collection of per-channel weights W by a gating mechanism consisting of two fully-connected layers and a sigmoid activation. Scale the multiplication between weights W and feature map U, and output the sum of the result and the input of SE-resblock. We use summation for skip connection.

TABLE I. Comparison with state-of-the-art using *Avg.Error (lower is better)* at different maximum cut-off depths on the MVSEC. The best results are in bold and the second-best results are underlined. Our method performs better than all baselines in both day and night scenes. Specifically, the method denoted with [S] indicates the network is trained on synthetic dataset DENSE.

Sequence	Distance	Frame based			Event based				
		MonoDepth	Megadepth	Megadepth+	Zhu et al.	E2Depth[S]	Ours[S]	E2Depth	Ours
Outdoor day1	10m	3.44	2.37	3.37	2.72	4.60	4.37	<u>1.85</u>	**1.50**
	20m	7.02	4.06	5.65	3.84	5.66	5.22	<u>2.64</u>	**2.39**
	30m	10.03	5.38	7.29	4.40	6.10	5.66	<u>3.13</u>	**2.91**
Outdoor night1	10m	3.49	2.54	<u>2.40</u>	3.13	10.36	5.34	3.38	**2.16**
	20m	6.33	4.15	4.20	4.02	12.97	6.68	<u>3.82</u>	**2.91**
	30m	9.31	5.60	5.80	4.89	13.64	7.23	<u>4.46</u>	**3.43**

TABLE II. Detailed comparison with E2Depth on MVSEC dataset. ↓ indicates lower is better and ↑ higher is better. Best results are shown in bold. Our method performs better than E2Depth in almost all metrics in both day and night environments.

Method	Sequence	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	SI log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
E2Depth	Outdoor day1	0.346	**0.516**	8.564	0.421	0.172	0.567	0.772	0.876
Ours		**0.319**	0.553	**8.333**	**0.389**	**0.144**	**0.600**	**0.799**	**0.897**
E2Depth	Outdoor night1	0.591	2.121	11.210	0.646	0.374	0.408	0.615	0.754
Ours		**0.428**	**1.781**	**8.869**	**0.467**	**0.204**	**0.529**	**0.725**	**0.849**

gradient matching loss is defined as below:

$$L_{k,grad} = \frac{1}{n}\sum_{s}\sum_{u}|\nabla_x R_k^s(u)| - |\nabla_y R_k^s(u)| \tag{5}$$

where $R_k^s(u)$ refers to the residual at scale s. The ∇_x and ∇_y denote the calculation of horizontal and vertical gradients using the Sobel operator. The local smoothness loss L_{smooth} calculates the difference in depth estimates between each pixel and its neighboring pixels, which is defined as follows:

$$L_{smooth} = \sum_{u}\sum_{u'\in\mathcal{N}(u)}\rho(D_k(u) - D_k(u')) \tag{6}$$

where $\rho(x) = \sqrt{x^2+\sigma^2}$ denotes the Chabonnier loss function [23] and $\mathcal{N}(u)$ denotes the 4 neighbouring pixels around the pixel u.

IV. Experimental Evaluations

In this section, we implement the experiments on the real-world event dataset MVSEC [24] and synthetic event dataset DENSE [18] to prove the effectiveness of our proposed method. We compare Mixed-EF2DNet with image-based and event-based depth estimation algorithms. The classical metrics of depth estimation are used to assess the performance of different methods quantitatively. We present both qualitative and quantitative evaluation results.

A. Experiment Setup

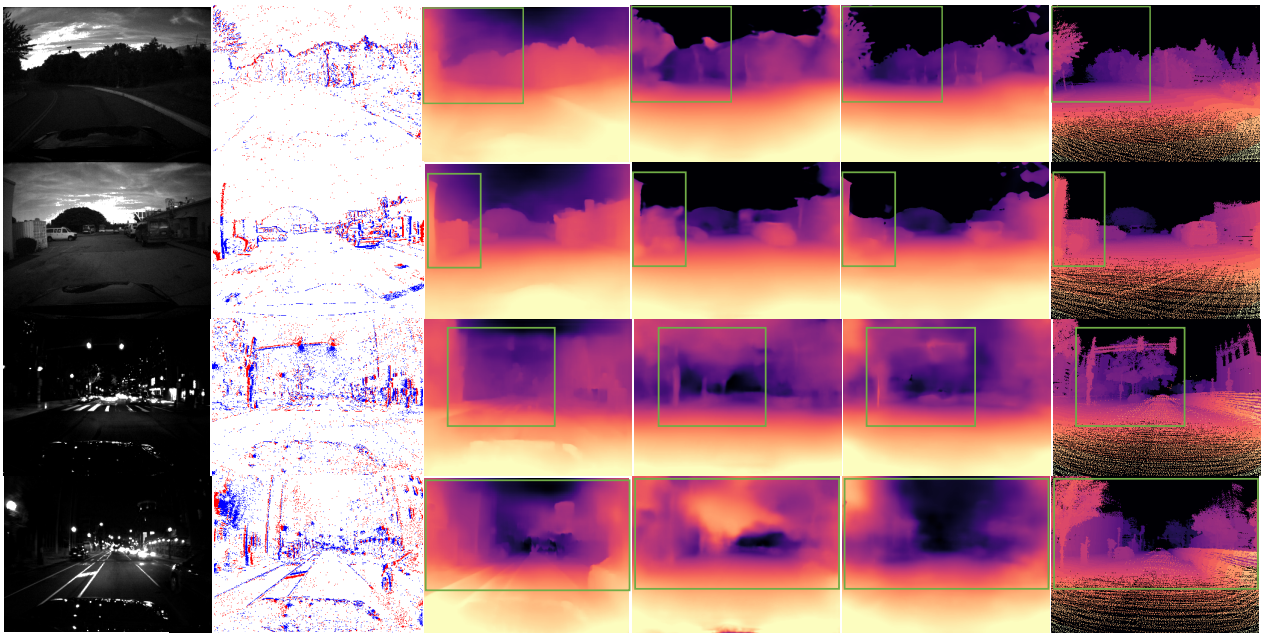
Datasets. The Multi-Vehicle Stereo Event Camera Dataset (MVSEC) is a popular real-world dataset recorded with mDAVIS346 sensors [24]. We use *outdoor day2* sequence for training, which contains 8523 samples of train split and 1826 samples of validation split. The representative *outdoor day1* and *outdoor night1* sequences of MVSEC are used for testing to verify the effectiveness of the depth estimation algorithm in the day and night environments.

The DENSE dataset with city road scenes is generated by event camera sensors based on ESIM [25] simulator. It contains 5000 training samples, 3000 validation samples, and 1000 testing samples. Compared with the real-world event dataset, the ground truth labels of the synthetic datasets have better quality with no missing values, which facilitates the training of supervised networks.

Implementation Details. We set the number of voxel grids bins B to 5, and the length of the input voxel grid list to 3. Mixed-EF2DNet is implemented using the PyTorch and optimized by the ADAM optimizer. The learning rate is set to 10^{-4}, and the batch size is set to 16. The weight factors λ_c and λ_s of the final depth map formulation are set to 1 and 0.5, respectively. The hyper-parameters $\lambda_1, \lambda_2, \lambda_3$ of objective function are set to 1.0, 0.25, 0.25, respectively.

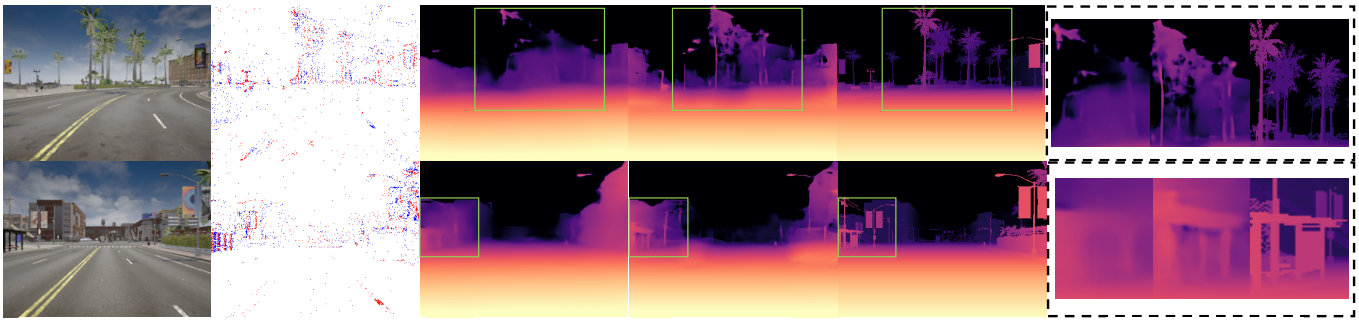
B. Evaluations on MVSEC Dataset

We first evaluate our proposed Mixed-EF2DNet on MVSEC Dataset to validate the effectiveness of our method in real-world environments. Following the prior work [18], we pre-train Mixed-EF2DNet with the first 1000 training



(a) Image Input (b) Event Input (c) MegaDepth (d) E2Depth (e) Ours (f) Ground Truth

Fig. 5. Qualitative comparison for MVSEC dataset. Each row presents a scene, with the scenes from *outdoorday1* on the top two rows and *outdoornight1* on the bottom two rows. In day environments, the predicted depth maps from our method avoid artifacts in the sky and have sharper objects boundaries than other baseline methods. In night environments, our method better estimates the detailed areas than other methods. Besides, it is less affected by noise event and predicts more robust depth maps compared to event-based baseline.



(a) Image Input (b) Event Input (c) E2Depth (d) Ours (e) Ground Truth (f) Scene Details

Fig. 6. Qualitative comparison for DENSE. Each row presents a scene. In (f), the details of E2Depth, ours, and ground truth are shown from left to right. Our proposed method provides a more complete and accurate depth estimation of objects, such as trees and bus stations.

samples of DENSE and then finetune the network by the combination train split of MVSEC and DENSE. We compare our proposed Mixed-EF2DNet with event-based depth estimation algorithms, including the method proposed by Zhu et al. [15] and E2Depth [18]. Besides, Mixed-EF2DNet also compares against well known image-based depth estimation algorithms, MonoDepth [8] [1] and MegaDepth [9], to verify the superiority of algorithms specifically designed for events.

Quantitative Evaluation. We quantified the performance of different methods using the average absolute error at depths of 10m, 20m, and 30m (*Avg.Error*; lower is better). The quantitative results are presented in Table I. MegaDepth+ indicates that events are reconstructed into images using E2VID [27] then taken by MegaDepth to estimate depth. The metrics values for MonoDepth and Zhu et al. method are taken from [15] and those for MegaDepth, MegaDepth+, and E2Depth from [18]. We observe that our method shows better performance than baseline methods in all metrics. Mixed-EF2DNet improves *Avg.Error* at all distances by an average of 11.8% in *outdoor day1* sequence and 19.0% in *outdoor night1* sequence. Note, night environments, where conventional images lose scene information and event streams have more noise due to low lighting, are more challenging to estimate depth than day environments.

We further compare our method in detail with E2Depth that performs best among all baseline methods using classical metrics of depth estimation [28], including (i) Absolute relative difference (Abs. Rel); (ii) Square relative difference (Sq. Rel); (iii) Root mean square difference (RMSE and RMSE log); (iv) Scale-invariant Error (SI log) and (v) Accuracy with threshold δ. As shown in Table II, our method performs better than E2Depth in seven of eight metrics measured for *outdoor day1* sequence. Besides, it achieves the best performance on all metrics for *outdoor night1* sequence, obtaining up to 45.4% improvement in *SI log* compared to E2Depth. The demonstrated results in *outdoor night1*

[1]MonoDepth performs accurately than MonoDepth2 [26] for the MVSEC dataset.

TABLE III. Comparison with state-of-the-art on the DENSE dataset. ↓ indicates lower is better. Our method improves the performance on five of six metrics, and achieves a score is approximate to E2Depth on *RMSE*.

Method	Avg. Error↓			Abs. Rel↓	Sq. Rel↓	RMSE↓
	10m	20m	30m			
E2Depth	0.610	1.450	2.420	0.220	0.279	**11.812**
Ours	**0.302**	**1.230**	**2.176**	**0.187**	**0.180**	11.873

sequence also show that Mixed-EF2DNet is robust when facing poor lighting environments with high noise levels.

Qualitative Evaluation. As shown in Fig. 5, we visualize the depth maps predicted by the image-based method MegaDepth, event-based method E2Depth, and Mixed-EF2DNet. For day environments, predicted depth maps from MegaDepth have artifacts in the sky. This is because the rich texture of clouds in images makes image-based methods challenging. However, the extreme distance and slow speed of the clouds result in slight relative motion from clouds to event cameras, leading to few events being triggered, thus avoiding interference from the sky for event-based methods. Compared to event-based method E2Depth, our Mixed-EF2DNet performs better at the depth prediction of object boundaries (e.g., the trees and buildings). For night environments, the dynamic range limits of conventional cameras are obvious. The predictions from MegaDepth might omit the scene information lost by images due to low lighting (e.g., the traffic light), and they have texture edges that break the continuity of depth (e.g., textures on the ground). E2Depth may misjudges noise as an object with a small depth value when there is much noise in the scene (e.g., scenes in the third column). In comparison, Mixed-EF2DNet, which incorporates contextual depth, suffers less from noise and produces depth predictions with spatially coherent. Besides, it performs better in the detailed areas than other methods.

C. Evaluations on DENSE Dataset

We compare our proposed Mixed-EF2DNet with the state-of-the-art event-based monocular depth estimation method, E2Depth, on the synthetic DENSE dataset. We train Mixed-EF2DNet on the training split of DENSE. The performance of methods is evaluated with main metrics. The values for E2Depth here are taken from [18].

Quantitative Evaluation. We first evaluate methods on the testing split of the DENSE dataset (as shown in Table III). Mixed-EF2DNet improves over E2Depth on *Avg.Error* of all estimated distances, by an average 25.3% relative gain. Besides, our method also improves *Abs.Rel* by 15.0% and *Sq.Rel* by 35.5%. The value of our method on *RMSE* is approximate to the value of E2Depth.

We further evaluate the models trained on DENSE in the testing split of MVSEC to investigate the generalization of methods. The comparison of *Avg.Error* are shown in Table I (the columns denoted with [S]). In both *outdoorday*1 and *outdoornigh*1 sequences, Mixed-EF2DNet performs better than E2Depth in all metrics by large margins, indicating that our method has better generalization across different datasets.

Qualitative Evaluation. Fig. 6 shows the qualitative comparison of the DENSE dataset. Each row corresponds to a scene from the testing split of the DENSE. We noticed that Mixed-EF2DNet has a more complete depth estimation of objects than the baseline method. Besides, our method estimates the depth of the scene more accurately and outputs depth maps with sharper surface boundaries aligning to local details of objects (e.g., the bus station and trees).

TABLE IV. Ablation study on the MVSEC dataset. Evaluate different methods using *Avg.Error*. The best results are in bold. Our proposed method offers the best in terms of all listed metrics.

Method	Outdoor day1			Outdoor night1		
	10m	20m	30m	10m	20m	30m
Mixed-EF2DNet	**1.50**	**2.39**	**2.91**	**2.16**	**2.91**	**3.43**
Mixed-EF2DNet_np	1.61	2.56	<u>3.01</u>	2.41	3.06	<u>3.50</u>
Mixed-EF2DNet_nl	<u>1.58</u>	<u>2.52</u>	3.05	<u>2.24</u>	<u>3.01</u>	3.57
Mixed-EF2DNet_npl	1.79	2.60	3.13	2.96	3.56	4.23
Mixed-EF2DNet_nspl	1.72	2.68	3.17	2.98	3.96	4.52
E2Depth	1.85	2.64	3.13	3.38	3.82	4.46

D. Ablation Study

To further demonstrate the effectiveness of the proposed method, we conduct an ablation study on the MVSEC dataset. Table IV presents the results of our analysis. Mixed-EF2DNet_np and Mixed-EF2DNet_nl denotes Mixed-EF2DNet without optical flow compensation and local smoothness loss respectively, and Mixed-EF2DNet_npl denotes Mixed-EF2DNet without both. Mixed-EF2DNet_nspl replaces SE-reblocks in Mixed-EF2DNet_npl with normal residual blocks. The training data and strategy are the same as Mixed-EF2DNet in Section 4.2.

As shown in Table IV, compared to Mixed-EF2DNet_npl, Mixed-EF2DNet_np improves *Avg.Error* by an average of 10.9%, and Mixed-EF2DNet_nl improves by an average of 12.1%, indicating that the introduction of optical flow or local smoothness loss is valid for depth estimation. Mixed-EF2DNet_npl indicates an improvement over Mixed-EF2DNet_nspl in most metrics owing to the introduction of SE-resblocks. Besides, it achieves better performance on almost all distances over E2Depth. Compared with other methods, Mixed-EF2DNet achieved the best score among all metrics. It improves over Mixed-EF2DNet_nspl by an average of 18.3%, demonstrating the effectiveness of the combination of key components.

V. Conclusion

This paper presents a novel method for dense depth estimation, Mixed-EF2DNet, based on a monocular event stream with no additional sensors. By compensating the input with optical flow and fusing the predicted contextual depth maps, the proposed method accurately estimates depth for each pixel even in challenging environments. The results on real-world and synthetic datasets have shown that our method demonstrates better performance when compared with state-of-the-art methods. Besides, our method also shows fair good generalization capability in across-data tasks. In the future, we would like to extend our proposed architecture to multi-tasks such as depth prediction and ego-motion estimation. We also consider trying it in event-based visual odometry and visual tracking.

REFERENCES

[1] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 $\times$ 128 120 db 15 μs latency asynchronous temporal contrast vision sensor," *IEEE journal of solid-state circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[2] T. Stoffregen, G. Gallego, T. Drummond, L. Kleeman, and D. Scaramuzza, "Event-based motion segmentation by motion compensation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7244–7253.

[3] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 994–1001, 2018.

[4] T. Laidlow, J. Czarnowski, and S. Leutenegger, "Deepfusion: real-time dense 3d reconstruction for monocular slam using single-view depth and gradient predictions," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4068–4074.

[5] P. R. Palafox, J. Betz, F. Nobis, K. Riedl, and M. Lienkamp, "Semanticdepth: Fusing semantic segmentation and monocular depth estimation for enabling autonomous driving in roads without lane lines," *Sensors*, vol. 19, no. 14, p. 3224, 2019.

[6] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[7] D. Falanga, K. Kleber, and D. Scaramuzza, "Dynamic obstacle avoidance for quadrotors with event cameras," *Science Robotics*, vol. 5, no. 40, 2020.

[8] 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*, 2017, pp. 270–279.

[9] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2041–2050.

[10] X. Nie, D. Shi, R. Li, Z. Liu, and X. Chen, "Uncertainty-aware self-improving framework for depth estimation," *IEEE Robotics and Automation Letters*, vol. 7, no. 1, pp. 41–48, 2021.

[11] H. Kim, S. Leutenegger, and A. J. Davison, "Real-time 3d reconstruction and 6-dof tracking with an event camera," in *European Conference on Computer Vision*. Springer, 2016, pp. 349–364.

[12] E. Piatkowska, J. Kogler, N. Belbachir, and M. Gelautz, "Improved cooperative stereo matching for dynamic vision sensors with ground truth evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 53–60.

[13] D. Zou, F. Shi, W. Liu, J. Li, Q. Wang, P.-K. Park, C.-W. Shi, Y. J. Roh, and H. E. Ryu, "Robust dense depth map estimation from sparse dvs stereos," in *British Mach. Vis. Conf.(BMVC)*, vol. 1, 2017.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*. Springer, 2020, pp. 402–419.

[20] 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.

[21] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[23] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proceedings of 1st International Conference on Image Processing*, vol. 2. IEEE, 1994, pp. 168–172.

[24] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3d perception," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2032–2039, 2018.

[25] H. Rebecq, D. Gehrig, and D. Scaramuzza, "Esim: an open event camera simulator," in *Conference on robot learning*. PMLR, 2018, pp. 969–982.

[26] 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.

[27] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[28] 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.