2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 1-5, 2023. Detroit, USA

Self-supervised Event-based Monocular Depth Estimation using Cross-modal Consistency

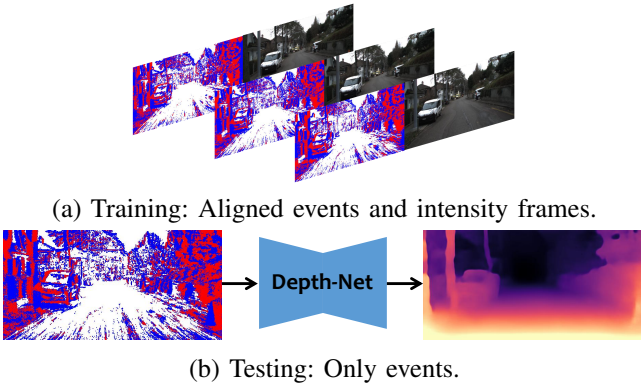
Junyu Zhu[1], Lina Liu[1], Bofeng Jiang[1], Feng Wen[2], Hongbo Zhang[2], Wanlong Li[2*] and Yong Liu[1*]

***Abstract*— An event camera is a novel vision sensor that can capture per-pixel brightness changes and output a stream of asynchronous "events". It has advantages over conventional cameras in those scenes with high-speed motions and challenging lighting conditions because of the high temporal resolution, high dynamic range, low bandwidth, low power consumption, and no motion blur. Therefore, several supervised monocular depth estimation from events is proposed to address scenes difficult for conventional cameras. However, depth annotation is costly and time-consuming. In this paper, to lower the annotation cost, we propose a self-supervised event-based monocular depth estimation framework named EMoDepth. EMoDepth constrains the training process using the cross-modal consistency from intensity frames that are aligned with events in the pixel coordinate. Moreover, in inference, only events are used for monocular depth prediction. Additionally, we design a multi-scale skip-connection architecture to effectively fuse features for depth estimation while maintaining high inference speed. Experiments on MVSEC and DSEC datasets demonstrate that our contributions are effective and that the accuracy can outperform existing supervised event-based and unsupervised frame-based methods.**

I. Introduction

Unlike conventional cameras, instead of capturing intensity frames at a fixed rate, event cameras only report brightness changes at the pixel level once they occur. The output of an event camera is a stream of asynchronous events in the format of (u_i, t_i, p_i) that encode the pixel location $u_i = (x_i, y_i)$, the time t_i, and the polarity p_i that denotes the sign of the brightness change that exceeds a threshold of $\pm C$. Such sensors have several advantages, e.g., very high dynamic range(140 dB vs. 60dB of conventional cameras), high temporal resolution and low latency(both in the order of microseconds), no motion blur, and low power consumption. These advantages give event cameras great potential for machine vision applications in challenging scenes with high-speed motions and high dynamic range. These years, event cameras are attracting the attention of computer vision researchers in various fields, including video reconstruction [1], visual odometry [2], optical flow estimation [3] and depth estimation [4], [5].

As a task of predicting a dense depth map from a single image, monocular depth estimation is an important and challenging field in computer vision. It helps computers understand the 3D structure of a scene. Thus, it can be applied in various fields such as autonomous driving, augmented reality, and 3D modeling. In the past few years, many monocular depth estimation methods that are based on conventional cameras have been proposed, including supervised methods [6], [7], self-supervised methods [8] and semi-supervised methods [9], [10]. Monocular depth estimation using event cameras is relatively less concerned [11], [12]. However, it is worth excavating the potential capacity of event cameras on depth estimation tasks due to their advantages in those application scenes which need high-frequency depth maps and adaptation to illumination changes.



(a) Training: Aligned events and intensity frames.

(b) Testing: Only events.

Fig. 1. **Method overview.** During the training(a), Pose-Net and Depth-Net are trained with aligned events and intensity frames. During the testing(b), Depth-Net estimates monocular depth map only from events.

Existing supervised event-based monocular depth estimation methods are fed with events [4] or a combination of events and frames [5] and trained with supervisory from ground truth depth maps collected by extra distance sensors. Expensive annotation cost is the obvious shortcoming of supervised methods. And existing unsupervised event-based monocular depth estimation methods usually depend on self-supervisory from event images deblurring [11], [12] or photo-consistency between the adjacent event images [13]. However, event data are sparse, so supervisory from events are not dense enough to constraint networks. Furthermore, matching pixels on adjacent event images is difficult because events on corresponding pixels can be very different.

We propose a framework that consists of Depth-Net and Pose-Net. Depth-Net and Pose-Net are jointly trained with the supervisory signal from the cross-modal consistency of intensity frames that are aligned with events in the pixel coordinate. As shown in Fig. 1, the intensity frames are only used for training, and at the testing time, our Depth-Net infers monocular depth maps using events as inputs only.

[1]Junyu Zhu, Lina Liu, Bofeng Jiang and Yong Liu are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China. email:{junyuzhu, linaliu, 22160068}@zju.edu.cn, yongliu@iipc.zju.edu.cn.

[2]Wanlong Li, Feng Wen and Hongbo Zhang are with Noah's Ark Lab, Huawei Technologies, Beijing, China. email:{liwanlong, wenfeng3, zhanghongbo888}@huawei.com.

*Corresponding authors: Wanlong Li and Yong Liu.

2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-9190-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IROS55552.2023.10342434

Besides, the decoder of conventional U-Net-based Depth-Net usually fuses upsampling feature maps and original feature maps from the encoder to get multi-scale feature maps for estimating multi-scale depth maps. We design a multi-scale skip-connection architecture based on the finding that Depth-Net can perform better when the above fusion process includes those lower-level feature maps.

To summarize, our contributions are the following:

- We find the photo-consistency of chronological events is weak for self-supervised learning.
- We propose a self-supervised framework that exploits cross-modal consistency from intensity frames for event-based monocular depth estimation.
- We introduce a multi-scale skip-connection architecture to effectively fuse features for depth estimation.
- We demonstrate that our framework can outperform existing supervised event-based methods and unsupervised frame-based methods.

II. Related works

A. Supervised Monocular Depth Estimation

Directly using the ground truth depth maps to supervise the training of networks is the most intuitive way to learn the monocular depth. [14] is the first to propose a supervised learning-based method to learn dense monocular depth. They consider monocular depth estimation as a regression task, and later works have followed such an idea for several years. However, [15] found that when depth estimation is regarded as a classification task, the depth estimation network can achieve better performance. And in these years, with the emergence of vision transformer [16], the performance of various visual tasks [17], [18], including monocular depth estimation [6], [7], has been significantly improved. Supervised monocular depth estimation models usually need a huge amount of parameters to achieve high accuracy, e.g., AdaBins [6] has $78M$ parameters, and NeWCRFs [7] has $270M$ parameters. Furthermore, obtaining GT depth maps at a high cost also limits the application of supervised methods.

B. Self-supervised Monocular Depth Estimation

To release the burden of collecting ground truth depth maps using expensive distance sensors, e.g., LiDAR, self-supervised monocular depth estimation is proposed to learn monocular depth by exploiting photometric consistency between stereo pairs or monocular sequences. [19] is one of the earliest works using stereo pairs to learn monocular depth in a self-supervised manner, and [20] introduced a left-right consistency loss to produce results comparable to early supervised methods. [21] extended such self-supervised framework to monocular sequences using an extra network to predict relative poses between adjacent frames. As a milestone in the self-supervised monocular depth estimation field, [8] helped the accuracy reach new heights by introducing an auto-masking technique and minimum reprojection loss to handle moving objects and occluded areas. To further enhance the accuracy, later works mainly focus on designing more complex network architectures [22], [23], using extra semantic constraints [24], [25]. Recently, several works [26], [27] further enhanced the self-supervised method by adopting novel data augmentation approaches to force networks to focus on the key information of images. Also, some works paid attention to the scale ambiguity of self-supervised methods, which are based on monocular sequences and solved the such problem by exploiting prior camera height [28], prior object size [29] and linear velocity measurement [30].

C. Event-based Depth Estimation

Depth estimation based on conventional cameras is prone to be affected by challenging illumination and limited by a fixed frame rate. Therefore, in those scenes with challenging lighting conditions and high-speed motions, depth estimation with event cameras is promising to get more reliable and higher-frequency depth maps.

Some methods estimate monocular depth using non-learning or unsupervised approaches. [31] presented a non-learning semi-dense depth estimation approach for a stereo event camera moving in a static scene. Their method first optimizes an energy function based on the spatiotemporal consistency of events triggered across both stereo image planes simultaneously, then uses a probabilistic fusion strategy to improve density and certainty of estimation. [11] proposed a method to estimate monocular depth from events by maximizing the variance of an image of warped events under the limiting assumptions that the camera pose is known and the scene is static. Inspired by [11], [12] trained networks in a self-supervised manner to predict optical flow, ego-motion, and depths with the goal of deblurring the event images.

Also, there are supervised methods for different settings with event cameras. For leveraging the temporal information, [4] uses a recurrent network for depth estimation from events. [5] extended this recurrent network for combining asynchronous events and synchronous frames to overcome their demerits and utilize their merits. [32] proposed a depth completion network that generates depth maps from sparse events and lidar point clouds.

III. Method

A. Method Overview

In this paper, we aim to train a network that can predict dense monocular depth from a continuous stream of events without ground truth depth maps. To this end, we propose a framework that consists of a Pose-Net and Depth-Net. The input of Depth-Net is event spatiotemporal voxel, which we describe in Sec. III-B, and the input of Pose-Net is adjacent intensity frames aligned with events. Our framework exploits cross-modal consistency loss, as described in III-E from adjacent intensity frames. Additionally, we design a multi-scale skip-connection architecture introduced in III-F for better performance of Depth-Net. The pipeline of our framework is shown in Fig. 2. For the training, we split events into subsequent non-overlapping windows of events $\varepsilon_k = \{e_i\}_{i=0}^{N-1}$ each spanning a fixed interval $\Delta T = t_{N-1}^k - t_0^k$ where t_{N-1}^k is aligned with the time of the intensity frame I_k and ε_k is then converted to spatiotemporal voxel E_k. After

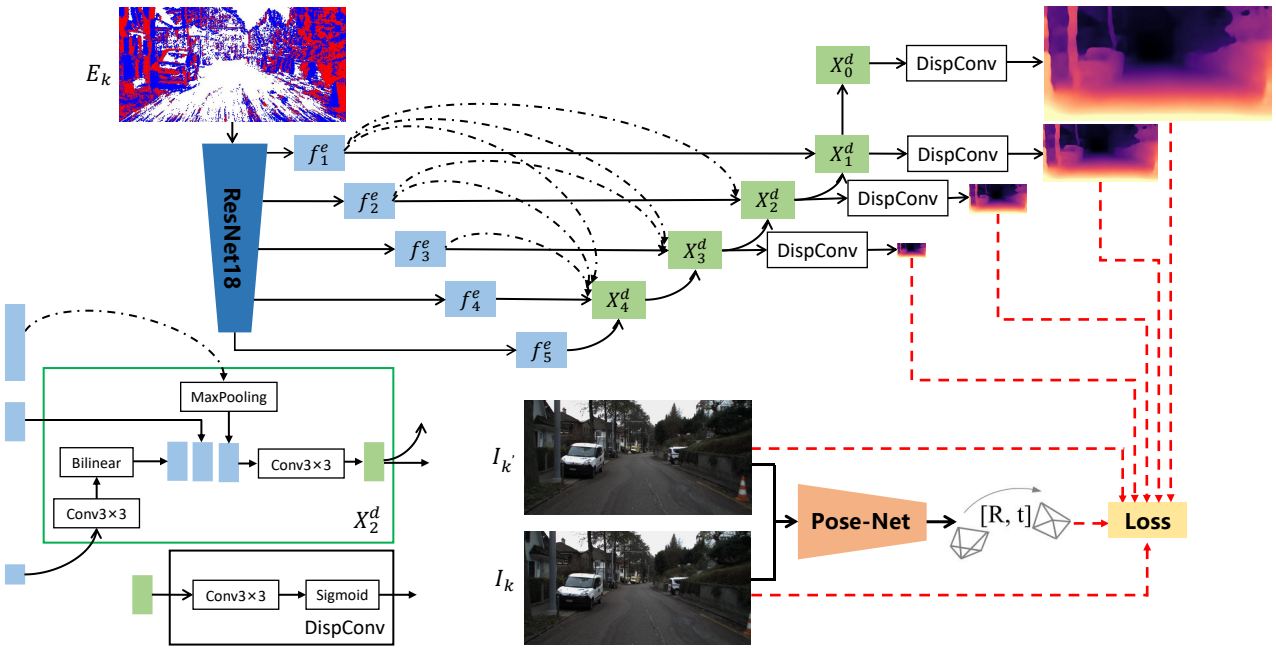


Fig. 2. **Framework illustration.** Our Depth-Net uses ResNet-18 as the encoder and uses several decoder nodes with multi-scale skip-connection architecture as the decoder. Multi-scale features f_i^e are encoded by ResNet-18 from event voxel grid E_k. These features are fused by decoder nodes x_i^d that includes our proposed multi-scale skip-connection architecture, and then outputs of decoder nodes are converted to disparity maps by $DispConv$ blocks. At the same time, the responding intensity frame and adjacent intensity frame are fed to a Pose-Net to predict relative pose $[R|t]$. Finally, the cross-modal consistency loss is computed using multi-scale event-based depth maps, relative pose, and intensity frames. After training, Depth-Net can estimate high-frequency monocular depth only from events.

training, Depth-Net can infer high-frequency dense depth maps only from events within a fixed interval.

B. Event Representation

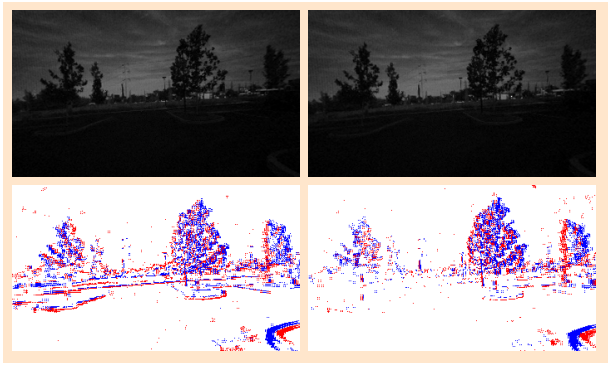
The output of an event camera is a stream of asynchronous events, and each event $e_i = (u_i, t_i.p_i)$ records the pixel location u_i, time t_i and polarity p_i of per-pixel brightness change. To feed a batch of events within the time window ΔT to networks, we convert the events to a tensor-like format E_k with a fixed dimension. There are several methods [33], [34], [35], [36], [37] proposed to represent events. In this paper, for a fair comparison with the majority of learning-based monocular depth estimation methods for events, we represent events as a spatiotemporal voxel grid with a dimension of $B \times H \times W$ that expresses events within the time window ΔT as B temporal bins. The converting process can be formulated as follows:

$$E_k(u_k, t_n) = \sum_{u_i = u_k} p_i max(0, 1 - |t_n - t_i^*|) \qquad (1)$$

where u_k denotes pixel coordinate on the image plane of $H \times W$, t_n belonging to $[0, B-1]$ denotes order number of temporal bins, $t_i^* = \frac{B-1}{\Delta T}(t_i - t_0)$ is the normalized timestamp of event e_i and u_i is the pixel location of event e_i. In this paper, we represent events within $\Delta T = 50ms$ as a spatiotemporal voxel grid with $B = 5$.

C. Cross-modal Consistency Loss

For constraining Pose-Net and Depth-Net in a self-supervised manner, the very intuitive idea is to exploit the photoconsistency between adjacent event spatiotemporal voxels. However, we found the photoconsistency is unsuitable for events in the adjacent time windows (see Fig. 3) because brightness changes on corresponding pixels can have contrary polarity and very different normalized timestamps that are closely related to camera motion. Moreover, the event spatiotemporal voxel is not dense enough (<15% in the MVSEC dataset) to provide networks with intensive self-supervisory signals.



(a) (b)

Fig. 3. **Adjacent event spatiotemporal voxels visualization.** The first row shows grayscale intensity frames from MVSEC, and the second row shows aligned events. We sum the voxel along the channel axis, then use blue and red pixels to represent negative and positive results. (a) and (b) are respectively from sample 125 and sample 126 of sequence $outdoor\_day1$. As shown in the figure, corresponding pixels can have very different events. So, the photoconsistency of event spatiotemporal voxel is weak.

As a solution, we choose to get consistency signals from other modals, e.g. intensity frames that are aligned with events in the pixel coordinate. With the depth estimation D_k from

an event spatiotemporal voxel E_k and the pose estimation $T_{k\to k'}$ between the corresponding intensity frame I_k and the adjacent intensity frame $I_{k'}$, a synthesized intensity frame $I_k^{k'}$ can be reconstructed from frame $I_{k'}$ with depth D_k and pose $T_{k\to k'}$ using inverse-warping [21]. Then, combining SSIM [38] loss and L1 loss, re-projection error $pe(I_k, I_k^{k'})$ can be computed as following:

$$\begin{aligned} pe(I_k, I_k^{k'}) =& \frac{\alpha}{2}(1 - SSIM(I_k, I_k^{k'})) \\ &+ (1-\alpha)\|I_k - I_k^{k'}\|_1 \end{aligned} \tag{2}$$

where SSIM is used to measure the structural similarity of images and computed over a 3×3 pixel window, and α is set to be 0.85. We follow the per-pixel minimum reprojection loss introduced by [8] to handle occlusion. The cross-modal consistency loss L_p can be formulated as:

$$L_{cc} = \min_{k'} pe(I_k, I_k^{k'}) \tag{3}$$

where $k' \in \{k-1, k+1\}$ thus two frames temporally adjacent to I_k are used as source frames.

D. Auto-Masking

To mask those pixels that remain the same due to a relatively static state and low texture, we apply the auto-masking technique proposed in [8]. The non-static mask M_{ns} is defined as:

$$M_{ns} = [\min_{k'} pe(I_k, I_k^{k'}) < \min_{k'} pe(I_k, I_{k'})] \tag{4}$$

where [] is the Iverson bracket.

E. Training Loss

We combine cross-modal consistency loss L_{cc} and non-static mask M_{ns} as $L = M_{ns}L_{cc}$, and average over pixels and scales to get final training loss L:

$$L = \frac{1}{s}\sum_{j}^{s}(\frac{1}{T}\sum_{i}^{T} M_{ns}^{j}(p_i)L_{cc}^{j}(p_i)) \tag{5}$$

where s denotes the number of scales, p_i demotes a pixel coordinate and T denotes the number of pixels of an image.

F. Multi-scale Skip-connection

U-Net-based architecture has become a common choice [8], [39] for designing the decoder of Depth-Net in the self-supervised monocular depth estimation field. As one of the core components of U-Net, skip-connection is used for recovering information lost in the downsampling process [23]. For an event spatiotemporal voxel, due to its natural sparsity (<15% in the MVSEC), recovering information lost becomes essential. Moreover, we think skip-connection from the encoder feature with the same level is insufficient. Inspired by [40], we propose a multi-scale skip-connection that connects with the feature of the same level and the lower-level features.

Let f_i^e denote a feature from the encoder, x_i^d denote the output of decoder node X_i^d. The x_i^d can be computed as:

$$x_i^d = \begin{cases} D(U(x_{i+1}^d)), & i = 0 \\ D([[f_i^e, [M(f_k^e)]_{k=1}^{i-1}], U(x_{i+1}^d)]), & i > 0 \end{cases} \tag{6}$$

where $U(\cdot)$ denotes an upsampling block consists of 3×3 convolution layer with ELU activation and bilinear interpolation. $M(\cdot)$ is a maxpooling layer to downsample lower-level features, $D(\cdot)$ is a feature fusion block made up of 3×3 convolution layer with ELU activation and $[\cdot]$ denotes the concatenation operation. The details of multi-scale skip-connection can be seen in Fig. 2.

With such designing, lower-level features can directly participate in the feature fusion process. Thus, information is less lost, and Depth-Net can achieve better performance.

IV. Experiments

In this section, we evaluate our proposed framework on the MVSEC dataset to present its qualitative and quantitative results and compare them with previous works. Also, ablation studies on the MVSEC dataset are conducted to demonstrate the effectiveness of our contributions. Further, to test our framework on higher resolution events and color intensity frames in more scenes, we evaluate the framework on the DSEC dataset, a very new event camera dataset.

A. Datasets

MVSEC. MVSEC[1] is the most commonly used dataset for the event-based depth estimation task. It was captured by a synchronized stereo pair event-based camera system carried on a handheld rig, flown by a hexacopter, driven on top of a car, and mounted on a motorcycle in different scenes and at various illumination levels.

DSEC. DSEC is a recently proposed dataset that offers data from a wide-baseline stereo setup of two color frame cameras with a resolution of 1080×1440 and two monochrome event cameras with a resolution of 480×640, and a lidar. DSEC is in driving scenarios and contains 53 sequences in different illumination conditions.

B. Inplementation Details

Network architectures. We use the same Pose-Net as previous works [8]. Moreover, we revise the first convolution layers of Depth-Net and Pose-Net to fit the channel number of events voxel and grayscale intensity frame. Considering different distributions of depth, the output σ of the Depth-Net is further constrained between 0.1 and 100 units for the MVSEC dataset and 0.1 and 60 units for the DSEC dataset with $D = 1/(a\sigma + b)$. For the MVSEC dataset, we take depth estimation of 4 scales for training, while only the maximum scale depth estimation is used for the DSEC dataset.

Data preprocessing of MVSEC. For the MVSEC dataset, we use the split proposed in [4]. More specifically, we train our networks on $outdoor\_day2$ sequence that is splited into 8523 training samples, 1826 validation samples, and 1826 testing samples. After training, we evaluate networks on other outdoor sequences. Considering static frames can disturb self-supervised training, we follow Zhou $et\ al.$'s [21] pre-processing and get 6817 training samples. To get a suitable input size, we improve resolution from the original 260×346

[1] https://daniilidis-group.github.io/mvsec/

TABLE I

QUANTITATIVE RESULTS ON THE MVSEC DATASET. COMPARISON OF THE AVERAGE ABSOLUTE DEPTH ERROR(IN METERS) AT DIFFERENT CUTOFF DISTANCE. $\mathbb{I}$ MEANS USING INSTENSITY FRAMES AS INPUT. $\mathbb{E}$ MEANS USING EVENTS AS INPUT. AND $\mathbb{I}+\mathbb{E}$ MEANS USING BOTH INTENSITY FRAMES AND EVENTS AS INPUT.

Dataset	Cutoff	Supervised hybrid	Supervised frame-based	Unsupervised frame-based	Supervised event-based	Unsupervised event-based	
		RAM-Net($\mathbb{I}+\mathbb{E}$) [5]	RAM-Net($\mathbb{I}$) [5]	Ours($\mathbb{I}$)	E2Depth [4]	Zhu et al. [12]	Ours($\mathbb{E}$)
outdoor day1	10m	**1.39**	1.74	1.54	1.85	2.72	1.40
	20m	2.17	2.55	2.23	2.64	3.84	**2.07**
	30m	2.76	3.07	2.71	3.13	4.40	**2.65**
outdoor night1	10m	2.50	2.72	3.24	3.38	3.13	**2.18**
	20m	3.19	3.35	3.74	3.82	4.02	**2.70**
	30m	3.82	3.99	4.60	4.46	4.89	**3.64**
outdoor night2	10m	**1.21**	1.36	3.16	1.67	2.19	2.06
	20m	**2.31**	2.42	3.65	2.63	3.15	2.76
	30m	**3.28**	3.47	4.24	3.58	3.92	3.42
outdoor night3	10m	**1.01**	1.20	3.09	1.42	2.86	2.09
	20m	2.34	2.44	3.55	**2.33**	4.46	2.82
	30m	3.43	3.64	4.20	**3.18**	5.05	3.52

TABLE II

QUANTITATIVE RESULTS ON THE DSEC DATASET. $\mathbb{I}$ MEANS USING INSTENSITY FRAMES AS INPUT AND $\mathbb{E}$ MEANS USING EVENTS AS INPUT. RESULTS IN LAST THREE COLUMNS DENOTE AVERAGE ABSOLUTE DEPTH ERRORS (IN METERS) AT DIFFERENT MAXIMUM CUT-OFF DEPTHS.

Method	Abs Rel↓	Sq Rel ↓	RMSE↓	RMSE log↓	SI log↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	$c = 10 \downarrow$	$c = 20 \downarrow$	$c = 30 \downarrow$
Ours($\mathbb{I}$)	0.166	1.346	5.934	0.229	0.050	0.749	0.934	0.980	1.042	1.942	2.622
Ours($\mathbb{E}$)	**0.142**	**1.152**	**5.258**	**0.202**	**0.043**	**0.819**	**0.948**	**0.983**	**0.960**	**1.550**	**2.205**

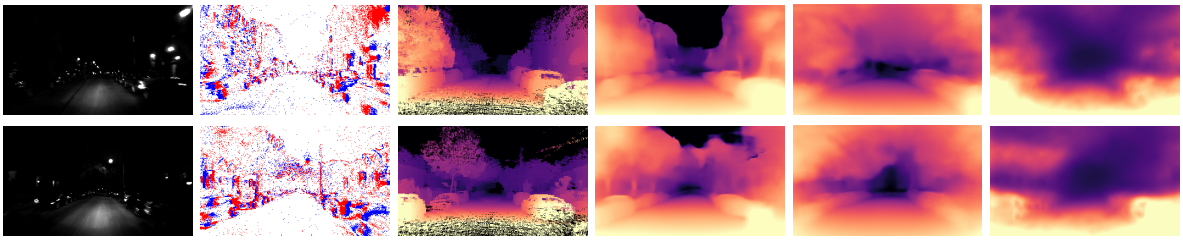


Fig. 4. **Qualitative results on the MVSEC dataset.** The qualitative result of Zhu et al. [12] is omitted because their code isn't publicly available. Our EMoDepth has a relatively more reasonable prediction on distant areas, e.g., trees and sky in the upper part of the image.

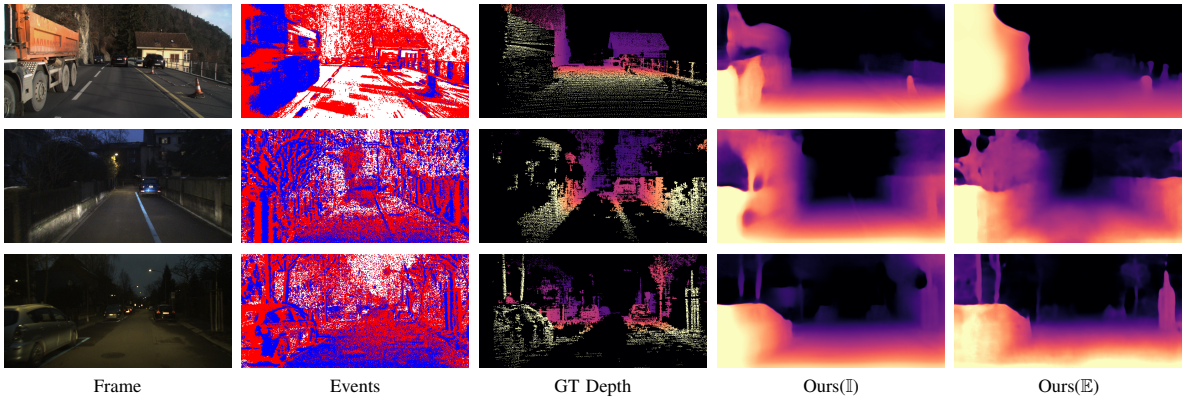


Fig. 5. **Qualitative results on the DSEC dataset.** $\mathbb{I}$ means using instensity frames as input and $\mathbb{E}$ means using events as input.

to larger 288×352 using zero padding. An evaluation is done on the upper middle areas with a resolution of 200×346.

Data preprocessing of DSEC. For the DSEC dataset, there are only optical flow benchmark and stereo matching benchmark on the official website[2] and the ground truth disparity maps of official testing sequences are reserved by the official server. For convenience, we split official training sequences into 28 for training and 13 for testing. The training set containing 16838 images has all images per training scene, and the testing set containing 1300 images has 100 images per testing scene. In the DSEC dataset, events and intensity frames are separately collected by different synchronized sensors. Following [41], we warp the rectified intensity frames to the event locations according to the calibration parameters provided. First, we get undistort and aligned event spatiotemporal voxels with intensity frames at a resolution of 480×640. Then, these data are center-cropped to 320×640 for cropping the car-hood.

[2]https://dsec.ifi.uzh.ch/

Data augmentation. We randomly flip the input images horizontally and apply color augmentations with a probability of 50%. For the color augmentation, we perform random brightness, contrast, saturation, and hue jitter by sampling uniform distributions in ranges of [0.8,1.2], [0.8,1.2], [0.8,1.2], [0.9,1.1], respectively. Note that color augmentation is only applied to the DSEC dataset.

Training setup. Our work is implemented in PyTorch on a single GTX1080Ti with 11GB memory. We train the networks for 10 epochs with Adam optimizer (β_1 = 0.9, β_2 = 0.999) and a batch size of 8. The initial learning rate is 1×10^{-4} for the first 8 epochs and 1×10^{-5} for the remaining.

C. Depth Estimation Results

We compare our framework performance with other methods on the MVSEC dataset. As in previous works, we report the average mean errors at maximum cutoff depths of 10m, 20m, and 30m. The quantitative and qualitative results are represented in Tab. I and Fig. 4, respectively. The quantitative results demonstrate that our EMoDepth outperforms unsupervised frame-based methods and even outperforms RAMNet($\mathbb{I}$) and E2Depth, which design recurrent architectures to leverage the temporal information and use additional synthetic data for better results. Due to the more practical constraints of proposed cross-modal consistency, EmoDepth performs better than Zhu et al. [12], which utilizes relatively weaker supervision by deblurring the event images. The qualitative results show that compared with E2Depth and RAMNet, our EMoDepth has a relatively more reasonable prediction on distant areas, e.g., trees and sky in the upper part of the image.

Also, we test our EMoDepth on the DSEC. Since there are no event-based monocular depth estimation methods reporting results on the DSEC dataset, we test our EMoDepth using frames($\mathbb{I}$) as input and using events($\mathbb{E}$) as input for comparison. The quantitative and qualitative results represented in Tab. II and Fig. 5 also show that when using the event as input, our EMoDepth can achieve better performance. Results also indicate that performance on the DSEC dataset is significantly better than on the MVSEC dataset. It can be explained as the benefits of denser events and color intensity frames.

D. Ablation studies

a) Choice of self-supervisory signal: We propose to exploit aligned intensity frames to form self-supervisory signals. To validate its effectiveness, we directly use the consistency of adjacent event spatiotemporal voxels to train the networks. As shown in Tab. III, the accuracy drops dramatically when the self-supervisory signals come from event spatiotemporal voxels. This is consistent with our analysis in III-C.

b) Effect of multi-scale skip-connection: We introduce multi-scale skip-connection to reduce information loss of sparse events. And the results in IV show that Depth-Net can achieve better performance when the multi-scale skip-connection is introduced when compared to the baseline (MonoDepth2).

TABLE III

ABLATION STUDIED OF DIFFERENT SELF-SUPERVISORY SIGNALS ON MVSEC DATASET

Dataset	Cutoff	Self-supervisory signal	
		Event spatiotemporal voxels	Intensity frames
outdoor day1	10m	3.90	**1.40**
	20m	3.79	**2.07**
	30m	4.89	**2.65**
outdoor night1	10m	5.55	**2.18**
	20m	4.57	**2.70**
	30m	5.72	**3.64**
outdoor night2	10m	5.76	**2.06**
	20m	4.48	**2.76**
	30m	5.26	**3.42**
outdoor night3	10m	5.87	**2.09**
	20m	4.39	**2.82**
	30m	5.33	**3.52**

TABLE IV

ABLATION STUDIED OF MULTI-SCALE SKIP-CONNECTION ON MVSEC DATASET.

Dataset	Cutoff	Baseline	+ multi-scale SC
outdoor day1	10m	1.48	**1.40 (↓0.08)**
	20m	2.22	**2.07 (↓0.15)**
	30m	2.74	**2.65 (↓0.09)**
outdoor night1	10m	2.55	**2.18 (↓0.37)**
	20m	3.06	**2.70 (↓0.36)**
	30m	3.95	**3.64 (↓0.31)**
outdoor night2	10m	2.47	**2.06 (↓0.41)**
	20m	3.03	**2.76 (↓0.27)**
	30m	3.64	**3.42 (↓0.22)**
outdoor night3	10m	2.44	**2.09 (↓0.35)**
	20m	2.95	**2.82 (↓0.13)**
	30m	3.60	**3.52 (↓0.08)**

c) Time consumption: We also test the inference time of E2Depth, RAM-Net, and our EMoDepth, on the MVSEC with a GTX1080TI GPU. The results in Tab. V show that the inference speed of EMoDepth is significantly faster than E2Depth and RAM-Net. And the introduction of multi-scale skip-connnection does not affect real-time performance much (from 4.67ms to 5.25ms) while improving performance.

TABLE V

TIME CONSUMPTION ON MVSEC DATASET USING DIFFERENT NETWORKS.

Network	Time consomption (ms)
E2Depth	24.22
RAM-Net	25.68
EMoDepth(w/o multi-scale SC)	**4.67**
EMoDepth	5.25

V. CONCLUSION

In this paper, we present a self-supervised framework named EMoDepth to learn monocular depth from events. Based on the observation that matching chronological events is challenging, we propose utilizing cross-model consistency from intensity frames. To fuse features more effective for improving performance, we design a multi-scale skip-connection architecture for EMoDepth. Extensive experiments on MVSEC and DSEC datasets show that EMoDepth can achieve state-of-the-art performance. In the feature work, we will explore exploiting events to remove moving objects to improve performance. Besides, a better event representation

method for retaining more spatiotemporal information of events while keeping a small data size is also a promising direction.

REFERENCES

[1] 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*, vol. 43, no. 6, pp. 1964–1980, 2019.

[2] J. Hidalgo-Carrió, G. Gallego, and D. Scaramuzza, "Event-aided direct sparse odometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5781–5790, 2022.

[3] M. Gehrig, M. Millhäusler, D. Gehrig, and D. Scaramuzza, "E-raft: Dense optical flow from event cameras," in *2021 International Conference on 3D Vision (3DV)*, pp. 197–206, IEEE, 2021.

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

[5] 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.

[6] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *CVPR*, 2020.

[7] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan, "Newcrfs: Neural window fully-connected crfs for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[8] 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 (ICCV)*, October 2019.

[9] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6647–6655, 2017.

[10] R. Ji, K. Li, Y. Wang, X. Sun, F. Guo, X. Guo, Y. Wu, F. Huang, and J. Luo, "Semi-supervised adversarial monocular depth estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2410–2422, 2019.

[11] G. Gallego, H. Rebecq, and D. Scaramuzza, "A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[12] 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 (CVPR)*, June 2019.

[13] C. Ye, A. Mitrokhin, C. Fermüller, J. A. Yorke, and Y. Aloimonos, "Unsupervised learning of dense optical flow, depth and egomotion with event-based sensors," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5831–5838, IEEE, 2020.

[14] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NeurIPS*, 2014.

[15] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *CVPR*, 2018.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference for Learning Representations (ICLR)*, 2021.

[17] J. Mei, M. Wang, Y. Lin, Y. Yuan, and Y. Liu, "Transvos: Video object segmentation with transformers," *arXiv preprint arXiv:2106.00588*, 2021.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020.

[19] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*, pp. 740–756, Springer, 2016.

[20] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017.

[21] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, 2017.

[22] F. Xue, G. Zhuo, Z. Huang, W. Fu, Z. Wu, and M. H. Ang, "Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2330–2337, IEEE, 2020.

[23] X. Lyu, L. Liu, M. Wang, X. Kong, L. Liu, Y. Liu, X. Chen, and Y. Yuan, "Hr-depth: High resolution self-supervised monocular depth estimation," in *AAAI*, 2021.

[24] S. Zhu, G. Brazil, and X. Liu, "The edge of depth: Explicit constraints between segmentation and depth," in *CVPR*, 2020.

[25] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, "Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance," in *European Conference on Computer Vision*, pp. 582–600, Springer, 2020.

[26] R. Peng, R. Wang, Y. Lai, L. Tang, and Y. Cai, "Excavating the potential capacity of self-supervised monocular depth estimation," in *ICCV*, 2021.

[27] H. Mu, H. Le, B. Yikai, R. Jian, X. Jin, and Y. Jian, "Ra-depth: Resolution adaptive self-supervised monocular depth estimation," in *ECCV*, 2022.

[28] A. Petrovai and S. Nedevschi, "Exploiting pseudo labels in a self-supervised learning framework for improved monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1578–1588, June 2022.

[29] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

[30] S. P. A. R. A. G. Vitor Guizilini Rares, Ambrus, "3d packing for self-supervised monocular depth estimation," in *CVPR*, 2020.

[31] 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)*, September 2018.

[32] 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.

[33] A. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Ev-flownet: Self-supervised optical flow estimation for event-based cameras," in *Proceedings of Robotics: Science and Systems*, (Pittsburgh, Pennsylvania), June 2018.

[34] R. Benosman, C. Clercq, X. Lagorce, S.-H. Ieng, and C. Bartolozzi, "Event-based visual flow," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 2, pp. 407–417, 2013.

[35] A. Zihao Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based optical flow using motion compensation," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

[36] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman, "Hats: Histograms of averaged time surfaces for robust event-based object classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1731–1740, 2018.

[37] D. Gehrig, A. Loquercio, K. G. Derpanis, and D. Scaramuzza, "End-to-end learning of representations for asynchronous event-based data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5633–5643, 2019.

[38] Z. Wang, "Image quality assessment : From error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004.

[39] H. Zhou, D. Greenwood, and S. Taylor, "Self-supervised monocular depth estimation with internal feature fusion," in *British Machine Vision Conference (BMVC)*, 2021.

[40] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1055–1059, IEEE, 2020.

[41] M. Mostafavi, K.-J. Yoon, and J. Choi, "Event-intensity stereo: Estimating depth by the best of both worlds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4258–4267, 2021.