2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

SRFNet: Monocular Depth Estimation with Fine-grained Structure via Spatial Reliability-oriented Fusion of Frames and Events

Tianbo Pan[1*], Zidong Cao[1*], *IEEE Student Member*, Lin Wang[1,2†], *IEEE Member*

***Abstract*— Monocular depth estimation is a crucial task to measure distance relative to a camera, which is important for applications, such as robot navigation and self-driving. Traditional frame-based methods suffer from performance drops due to the limited dynamic range and motion blur. Therefore, recent works leverage novel event cameras to complement or guide the frame modality via frame-event feature fusion. However, event streams exhibit spatial sparsity, leaving some areas unperceived, especially in regions with marginal light changes. Therefore, direct fusion methods, *e.g.*, RAMNet [6], often ignore the contribution of the most confident regions of each modality. This leads to structural ambiguity in the modality fusion process, thus degrading the depth estimation performance. In this paper, we propose a novel Spatial Reliability-oriented Fusion Network (SRFNet), that can estimate depth with fine-grained structure at both daytime and nighttime. Our method consists of two key technical components. Firstly, we propose an attention-based interactive fusion (AIF) module that applies spatial priors of events and frames as the initial masks and learns the consensus regions to guide the inter-modal feature fusion. The fused feature are then fed back to enhance the frame and event feature learning. Meanwhile, it utilizes an output head to generate a fused mask, which is iteratively updated for learning consensual spatial priors. Secondly, we propose the Reliability-oriented Depth Refinement (RDR) module to estimate dense depth with the fine-grained structure based on the fused features and masks. We evaluate the effectiveness of our method on the synthetic and real-world datasets, which shows that, even without pretraining, our method outperforms the prior methods, *e.g.*, RAMNet [6], especially in night scenes. Our project homepage: https://vlislab22.github.io/SRFNet.**

I. INTRODUCTION

Monocular depth estimation is a fundamental vision task with various applications, such as robotic navigation [28] and self-driving [8]. In the last decade, deep learning-based methods have shown significant performance gains for monocular depth estimation using the standard frame-based cameras [13], [9]. However, these methods encounter critical challenges, particularly in extreme, *e.g.*, high-speed motion and low-light scenes.

Recently, using novel event cameras to estimate depth has experienced a surge in popularity [39]. Event cameras are bio-inspired sensors that asynchronously detect the intensity changes of a scene and output event streams encoding time, pixel location, and polarity (*i.e.*, sign for the intensity changes). Event cameras show distinct advantages over RGB-based cameras, such as high dynamic range (HDR) and high temporal resolution. Accordingly, some event-based methods have been proposed to directly estimate depth from event data [10], [30]. Although these methods show superiority in extreme visual conditions, the intrinsic sparsity and unexpected noise of events often cause structural ambiguity of objects.

To complement the RGB cameras, recent works attempt to leverage event cameras as guidance for better depth estimation [6], [22]. These methods learn the modal-specific features independently and fuse RGB-event features without distinguishing the contribution of the most confident regions of each modality. This leads to structural ambiguity between the two modalities, which may hamper the depth estimation performance. Also, although the prior work RAMNet [6] considers the temporal information of event data, it overlooks the importance of spatial sparsity, which is crucial for structural information extraction. Intuitively, it remains a challenge to effectively harness the strengths of both modalities.

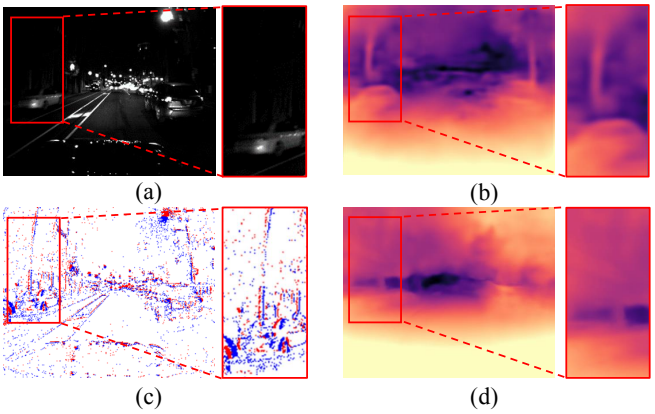


Fig. 1. (a) and (c) are the input intensity frame and events, respectively; (b) is our result while (d) is the result of RAMNet [6]. Within the red rectangle, we zoom in on a challenging scenario where a car and tree in the frame are exposed under poor lighting and show discontinuous edge information in the events. Our SRFNet excels in predicting dense depth with fine-grained structural details compared with RAMNet.

In this paper, we find that inter-modal cross-referencing can greatly benefit feature learning of each modality. Accordingly, we present a novel framework Spatial Reliability-oriented Fusion Network (**SRFNet**) that can estimate dense depth with fine-grained structure at both daytime and nighttime, as shown in Fig. 1(b). Our SRFNet consists of two key technical components (See Fig. 3). Firstly, we propose an Attention-based Interactive Fusion (AIF) module, including a series of Consensus Learning (CL) blocks to adaptively integrate the two modalities while employing the fused feature to enhance each modality's feature learning (Sec. III-B). Specifically, the AIF module initially applies spatial priors of event and frame as the initial masks to learn the consensus regions of two modalities and generate a fused feature through an attention-based layer. The fused features are then fed back to enhance the frame and event feature learning. Meanwhile, it leverages an output head to generate a fused mask for iterative spatial prior learning. That enable the spatial prior progressively to have more consensus information. Through the iterative fusion process of CL blocks, it is possible to effectively leverage the superiority of each modality.

Secondly, we propose a Reliability-oriented Depth Refine (RDR) module to estimate dense depth with the fine-grained structure based on the fused feature and masks. Concretely, we employ a

†Corresponding author, * Co-first authors.
[1] T. Pan is with AI Thrust, HKUST(GZ), Guangzhou, China, Email: tpan695@connect.hkust-gz.edu.cn
[1]Z. Cao is with AI Thrust, HKUST(GZ), Guangzhou, China, Email: caozidong1996@gmail.com
[1,2]L. Wang is with AI Thrust, HKUST(GZ) and Dept. of CSE, HKUST, China, Email: linwang@ust.hk

decoder together with a temporal learning layer to estimate a depth feature. Then, the fused masks from CL blocks are stacked and incorporated with the depth feature. The original depth feature is used to obtain the affinity map and coarse depth map with two individual output heads, respectively, while the incorporated depth feature is used to obtain the confidence map. Finally, the affinity map, coarse depth map, and confidence map are fed into the spatial propagation network (SPN) to generate the fine-grained depth map.

In summary, the main contributions of our paper are:

- We propose a novel SRFNet that can estimate dense depth with fine-grained structure at both daytime and nighttime.
- We propose the AIF module that interactively enhances two modalities' feature learning and the RDR module that further obtains fine-grained depth estimation.
- We conduct extensive experiments on both synthetic and real-world datasets. We compare our method with the state-of-the-art (SOTA) frame-based, event-based, and frame-event fusion methods. The experimental results reveal that *even without pre-training*, our SRFNet yields a **substantial improvement** over the SOTA methods, *e.g.*, RAMNet [6], and exhibits outstanding generalization capabilities, especially in night scenes.

II. Related Works

Monocular Depth Estimation. Frame-based monocular depth estimation has been studied in both images [1] and videos [34]. Previous approaches in this domain have primarily revolved around two key aspects: 1) Training strategy, including supervised learning [21], [5], unsupervised learning [34], and transfer learning [1]. 2) Backbone architectures, such as vision transformers (ViT) [21]. Although frame-based methods achieve excellent performance, their performance often drops greatly in challenging scenes, such as those with poor lighting conditions and high-speed motion.

Recently, event cameras have garnered substantial attention due to their ability to capture high dynamic range (HDR) scenes in extreme visual conditions [39], [10], [32], [36], [17]. This has led to increased research interests in event-frame fusion for monocular depth estimation. For example, EVEN [22] integrates multi-modal features to enhance the frame feature learning. RAMNet [6] introduces an RNN-based network to further utilize the asynchronous property of events for better fusion. However, these methods pay less attention to the spatial reliability of each modality during fusion, treating both informative and noisy regions equally. *Differently, our proposed SRFNet that can better distinguish the contributions of the most confident regions of each modality during fusion. It can estimate depth map with fine-grained structure, especially in nighttime scenes (See Fig. 1(b)).*

Event-Frame Fusion. The complementary nature of event and frame data has been extensively explored in various vision tasks, like object detection [31], [38], [40], [27], semantic segmentation [26], [37], [4], [15], and depth estimation [42], [11], [3], [6], [22]. The above mentioned fusion methods can be classified into two categories based on whether the interaction is considered.

The first type is fusion without interaction, where each modal feature is learned independently, and the fused feature is only utilized by the decoder. For instance, RAMNet [6] learns event and frame features asynchronously and separately. EVEN [22] and HDES [42] add multi-modal features and use the attention mechanism to further incorporate them, while [38] leverages the cross-modal attention to achieve multi-modal feature integration. Also, [27] proposes to concatenate synchronous multi-scale multi-modal features. The second type is fusion with interaction, which involves getting feedback from the fused result. Among the approaches, [31], [40], [18] explore the feedback paradigm, in which the learned fused feature is added with the modal-specific features to enhance the subsequent feature learning. [3] proposes to generate a cross-similarity feature and employs convolutions to merge the similarity feature with modal-specific features. However, these works fuse frame-event features without distinguishing the contribution of the most confident regions of each modality, which might result in structural ambiguity in the extreme scenes. *In contrast, our SRFNet applies spatial priors of events and frames and learns the consensus regions to guide the inter-modal feature fusion.*

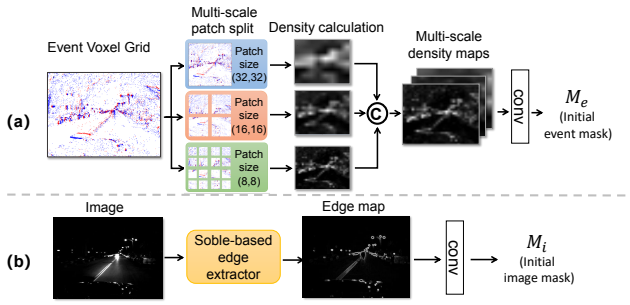


Fig. 2. Illustration of the initialization of modal-specific masks, M_i for the frame and M_e for the events.

Spatial Learning for Depth Estimation. Spatial Propagation Network (SPN) [16] aims to refine the coarse depth to achieve fine-grained details in the estimated depth map. For instance, CSPN++ [2] develops adaptable kernel sizes and iteration numbers to enhance the propagation process. The effectiveness of spatial learning is further boosted by incorporating deformable local neighbors [19], [33]. *Differently, we leverage the fused masks from the AIF module to refine the confidence map, thus giving more accurate guidance for the refinement process.*

III. Method

In this section, we first describe the data preprocessing, including the event representation and modality-specific mask initialization. We then outline the network architecture of our SRFNet framework, composed of the Attention-based Interactive Fusion (AIF) module and Reliability-oriented Depth Refine (RDR) module. We now describe the details.

A. Data preprocessing

Event Representation. An event $e = (x, y, t, p)$ is generated when the log intensity difference at a pixel location (x, y) and time t exceeds a threshold C. The polarity $p = \pm 1$ indicates the intensity change direction. Following [10], we accumulate events using non-overlapping event windows, spanning a fixed interval ΔT. The k_{th} event window is represented into a frame-like representation $\mathcal{E}_k = \{e_i\}_{i \in [0, \mathbf{N}-\mathbf{1}]}$, where N is number of event in the window and the fixed time interval $\Delta T = t^k_{N-1} - t^k_0$. encoded as spatio-temporal voxel grids V_k with dimensions $B \times H \times W$. Here, B denotes the number of temporal bins. Following [10], [6], we set B to 5 and normalize the voxel grid representation. The resulting volume V_k forms a frame-like tensor, making it compatible with processing using convolutional neural networks.

Modal-specific Mask Initialization. In order to get more effective information extraction as [25], we design learnable modality-specific masks that allow each modality to contribute its most confident regions. Instead of deriving masks from extracted features, we initialize them according to each modality's inherent properties and develop separate learning for iteratively updating modality-specific masks.

The mask serves as a confidence map to indicate the informative content of various regions. For initializing the event masks, each time bin of the voxel grid is split into multi-scale non-overlapping patches (*i.e.*, 8×8, 16×16, 32×32), as depicted in Fig. 2(a). The event density within each patch is calculated and then normalized by dividing it by the average density of all patches. We average the density maps from different time bins under the same patch size and concatenate the averaged density maps from multi-scale patches. These concatenated density maps are subsequently input into a 3×3 convolution to obtain an event mask M_e. Fig. 2(b) depicts the initialization of the frame mask, which is based on the assumption that regions with more edge information are more likely to have significant changes in depth [14], [20]. Our method employs the Sobel edge extractor [24] and feeds the edge map to convolution to learn to the frame mask M_i^k. These two masks are taken as the inputs to the SRFNet.

B. Network Architecture

As shown in Fig. 3, our SRFNet predicts dense log depth maps with fine-grained structure from preprocessed event voxel grids and frame sequence. For the k_{th} paired input: event voxel grids V^k, frames I^k, and event mask M_e^k and frame mask M_i^k, the first Attention-based Interactive Fusion(AIF) module generate fused features F^k and stacked fused masks $M_f^k u$, which contains consensual information learned from modalities. The fused features and masks are then passed to the second Reliability-oriented Depth Refine (RDR) module, which functions as a decoder, generating an initial coarse prediction and further refining its structural details.

Attention-based Interactively Fusion (AIF) Module. It receives preprocessed modal-specific masks and pyramid modal features from event and frame branches as input. It utilizes spatial priors to guide inter-modal fusion, promoting consensual fused feature learning and enhancing modal-specific feature learning.

Specifically, the event and frame branch extend the encoder design in [9]. The High-Resolution Network (HRNet) [29] is chosen as the encoder head due to its proven excellence in perceiving object layouts. F_e^k, F_i^k represents the event and frame feature extracted from the k_{th} frame and voxel grid I^k, V^k, respectively. Additionally, to align with the resolution of learned feature, the modal-specific masks M_e^k, M_i^k are downsampled using a convolution with a 3×3 kernel, a stride of 2, and padding of 1.

Our AIF module leverage multiple Consensus Learning (CL) blocks to facilitate the interactive fusion, as illustrated in Fig. 3. The CL block, depicted in Fig. 3(b), is an attention-based fusion block that generates fused features and masks. Prior to the combination, CL block emphasizes the modal-specific features $F_e, F_i \in R^{H\times W\times C}$, with their masks $M_e, M_i \in R^{H\times W\times 1}$, which represent spatial priors denoting pixel-wise reliability. The emphasized features are then concatenated and passed through an attention layer to learn a fused feature $F_{fused} \in R^{H\times W\times 2C}$. The fused feature F_{fused} is then split to gain feedback features $\overline{F}_e, \overline{F}_i$, and then add them to F_e^k, F_i^k, enhancing the subsequent event and frame feature learning. Meanwhile, the fused feature leverages an output head to generate a fused mask containing consensual information. Subsequently, the fused mask updates the modal-specific mask with a residual connection manner, as shown in Fig. 3(b). The event and frame masks are iteratively updated through the CL blocks to incorporate more consensual information. In the last CL block, the emphasized F_e^k, F_i^k are directly added and passed to attention layer to learn the final fused feature and mask. Ultimately, the AIF module stacks all masks generated by the CL blocks and then passes the fused features F^k and stacked masks M_{fused}^k to next module.

Reliability-oriented Depth Refinement (RDR) Module. Our Reliability-oriented Depth Refinement (RDR) module serves as an aggregation of decoder and refinement block. It enhances the accuracy of coarse predictions by incorporating a temporal layer that strengthens the fused feature with temporal correlations. Additionally, the RDR module optimizes the generation of affinities, significantly improving the performance of the NLSPN-based refinement block, resulting in fine-grained depth structures and reduced artifacts.

Inspired by [35], we optimize the decoder structure in [7] by adding a temporal layer near the output head, which learns temporal correlations. The temporal layer is replaceable, we use convolution Gated Recurrent Unit (convGRU) [23] in our SRFNet. The convGRU layer utilizes an updatable latent state S to capture temporal information. By integrating fused feature F^k and previous state S^{k-1} , it yields both the temporal feature $\widetilde{F}^k$ and current state S^k, as described in the following formulation:

$$(\widetilde{F}_{fused}^k, S^k) = convGRU(F_{fused}^k, S^{k-1}) \tag{1}$$

Subsequently, the $\widetilde{F}_{fused}^k$ is utilized to generate the coarse depth prediction via a depth output head, while S^k spreads the temporal information for $k+1_{th}$ prediction.

To further eliminate the ambiguity of structural details of depth map, the RDR module incorporates the Non-Local Spatial Propagation Network (NLSPN) [19] for refinement. NLSPN's advantage lies in its ability to search for neighbors beyond adjacent nodes, effectively capturing long-range spatial dependencies. The effectiveness of NLSPN greatly depends on the accuracy of estimated affinity. To ensure stability during propagation, normalizing the affinities is essential. Incorporating the confidence of the initial depth prediction during normalization helps mitigate the adverse effects of unreliable depth values during propagation, as demonstrated in the following equation:

$$w_{m,n}^{i,j} = c^{i,j} \cdot \tanh(\hat{w}_{m,n}^{i,j})/\gamma \tag{2}$$

$$c = f_\psi(\widetilde{F}_{fused}, M_{fused}) \tag{3}$$

$w_{m,n}^{i,j}$ and $\hat{w}_{m,n}^{i,j}$ denotes the normalized and estimated affinity of pixel (i,j) respectively. $c^{i,j} \in [0,1]$ represents it's confidence level, while γ denotes the learnable normalization parameter. The Eq. 2 ensures that the cumulative sum of normalized affinities within the neighborhood remains bounded by one. Moreover, Eq. 3 reveals the optimized generation process of confidence map c, where $f_\psi(\cdot)$ denotes the confidence head which leverages attention mechanism to integrate temporal feature $\widetilde{F}_{fused}$ and fused masks M_{fused}. The fused masks are essentially similar to the confidence map, as they reflect the consensual spatial reliability of modalities, providing extra guidance in the learning of Eq. 3.

Our RDR module significantly improves the effectiveness of NLSPN by enhancing the accuracy of estimated affinity and confidence. In affinity estimation, the learned feature $\widetilde{F}_{fused}$ effectively aggregates spatial-temporal information, benefiting from the global horizon of attention layer and the convGRU layer's exploration of temporal correlation. Regarding confidence estimation, the learned fused masks serve as informative extra features to enhance confidence map learning. Consequently, the RDR module with enhanced affinity enables spatial propagation to yield higher accuracy and fine-grained depth maps..

C. Loss Functions

Our SRFNet is trained in a supervised manner, where the depth ground truth is captured from the LiDAR. Following RAMNet [6],

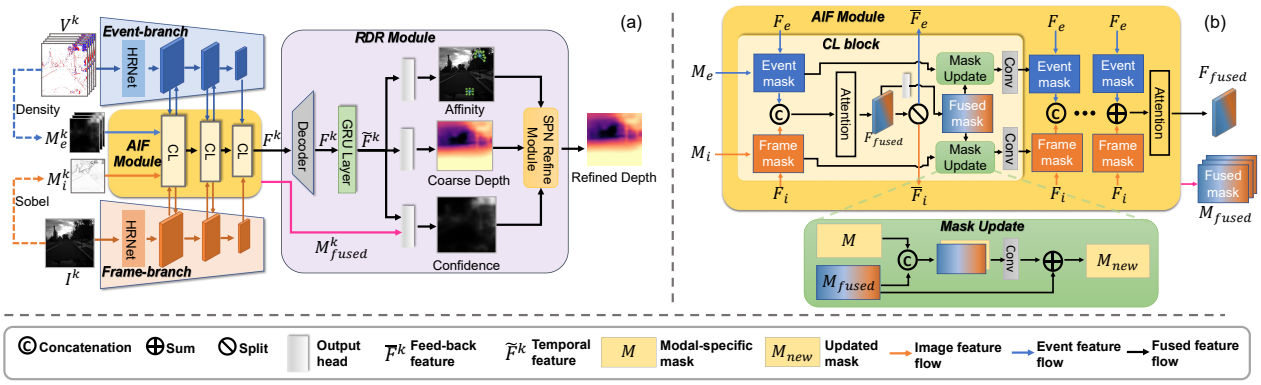


Fig. 3. (a) is the overview of our proposed framework, and (b) depicts the details of the AIF module.

the depth values are normalized into log formats $\bar{\hat{D}}_k \in [0,1]$, to facilitate predicting depths with large variations. The normalization can be formulated as follows:

$$\widehat{D}_k = \frac{1}{\alpha} \log \frac{\widehat{D}_{m,k}}{D_{\max}} + 1, \tag{4}$$

where α is a predefined parameter to map the closestvalid depth to 0. For MVSEC dataset, $\alpha = 3.7$ and $D_{max} = 80$m. Our loss comprises two components: MSE loss and multi-scale scale-invariant gradient matching loss, computed for valid ground truth labels and summed over a sequence of events and frames:

$$L_{total} = \sum_{k=0}^{L-1} L_{k,mse} + \lambda L_{k,grad} \tag{5}$$

where L denotes the length of sequence fed for training. And the λ is set to 0.25 in this paper. These losses are calculated based on the log depth difference $R_k = \widehat{D_k} - D_k$, where $\widehat{D_k}$ is the predicted depth and D_k is the ground truth depth. The MSE loss and gradient matching loss can be formulated as:

$$L_{k,mse} = \frac{1}{n} \sum_{\mathbf{u}} (R_k(\mathbf{u}))^2 \tag{6}$$

$$L_{k,grad} = \frac{1}{n} \sum_{s} \sum_{\mathbf{u}} |\nabla_x R_k^s(\mathbf{u})| + |\nabla_y R_k^s(\mathbf{u})| \tag{7}$$

Among them, u indexes the valid ground truth pixels and $R_k^s(u)$ refers to the residual at scale s. ∇_x and ∇_y compute the edges in the x and y direction respectively by using the Sobel operator.

IV. EXPERIMENTAL EVALUATION

A. Dataset and Implementation Details

We conduct experiments on two datasets, *i.e.*, MVSEC dataset and DENSE dataset. The MVSEC dataset is a real-world event dataset [41], while the DENSE dataset is a synthetic dataset [10].

MVSEC dataset: It is recorded with a pair of DAVIS cameras. The resolution of frames and events is 346×260. For supervised training, the depth ground truth is obtained via a LiDAR scanner. This dataset covers not only daytime and nighttime driving sequences, but also indoor sequences captured by a quadcopter. We choose the subset "outdoor day2" for training and "day1", "night1", "night2" and "night3" for testing. The frames are recorded with grayscale. The depth maps, daytime and nighttime frames are recorded at 20Hz, 10Hz, and 45 Hz, respectively.

Implementation Details: For event streams, we divide each voxel grid into 5 time bins. Similar with RAMNet [6], we set the length of input sequences with paired events and frames to 8. We utilize Adam optimizer [12]. The initial learning rates set for the AIF module and RDR module are 5e-6 and 1e-5, respectively. The batch size is 4, and the number of training epoches is 100. For data augmentation, we employ data normalization, randomly cropping, and horizontal flipping.

B. Evaluation on MVSEC and DENSE Datasets

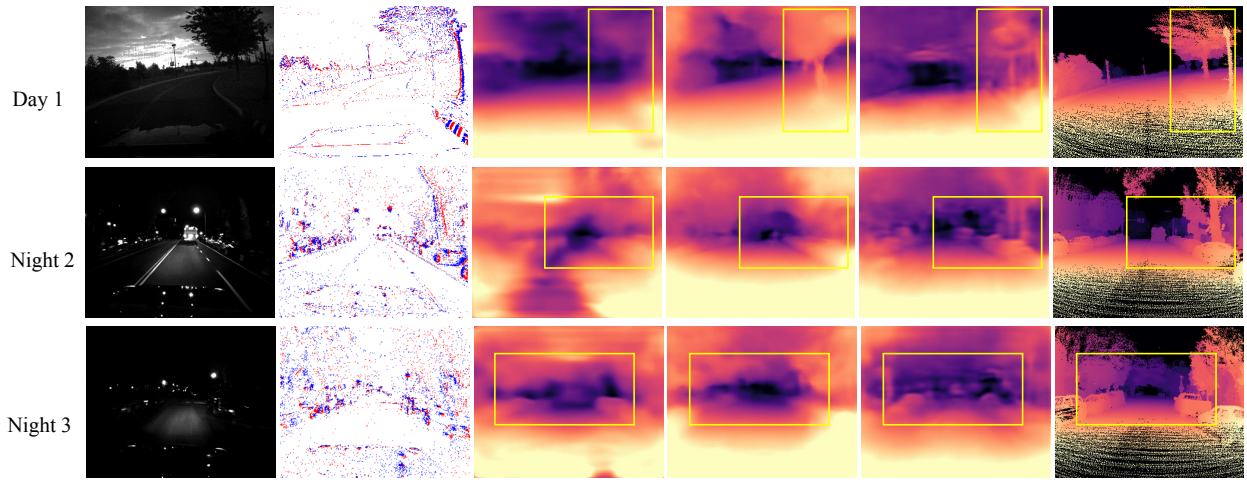
Previous methods [6], [10], [32] are pre-trained on synthetic datasets. Considering that the amount of these synthetic datasets is huge, we skip the pre-training to avoid tedious training process. For adequate evaluation, we not only compare with the methods based on frames [9], [13], events [10], but also compare with the methods combining frames with events [6]. In addition, to evaluate our SRFNet and compared methods in the same training setting, we re-train MonoDEVS [9] and RAMNet [6] on the MVSEC dataset.

Metrics. Similar with RAMNet [6], we choose average absolute depth errors at different cut-off depth distances (i.e., 10m, 20m and 30m) for comparison. The metrics we utilize include absolute relative error (Abs Rel), logarithmic mean squared error (RMSE log), scale invariant logarithmic error (SI log), and accuracy ($\delta < 1.25^n, n = 1,2,3$) for detailed comparison with RAMNet.

Quantitative Evaluation. As shown in Tab. I, we conduct comprehensive comparisons among different types of monocular depth estimation methods on various scenes. It can be seen that our method outperforms methods with single modality as input [9], [13], [10] in most scenes. Note that our method has lower performance compared with E2Depth in some scenes, due to its pre-training on synthetic datasets. Moreover, compared with RAMNet that combines frames with events, our method even outperforms it even with pre-training. For example, compared with RAMNet with pre-training, our method obtains 0.22 gain with 20m distance in the outdoor day1 dataset. *The results can also demonstrate the superior generalization of our method, as the training dataset we utilize contains no night scene.*

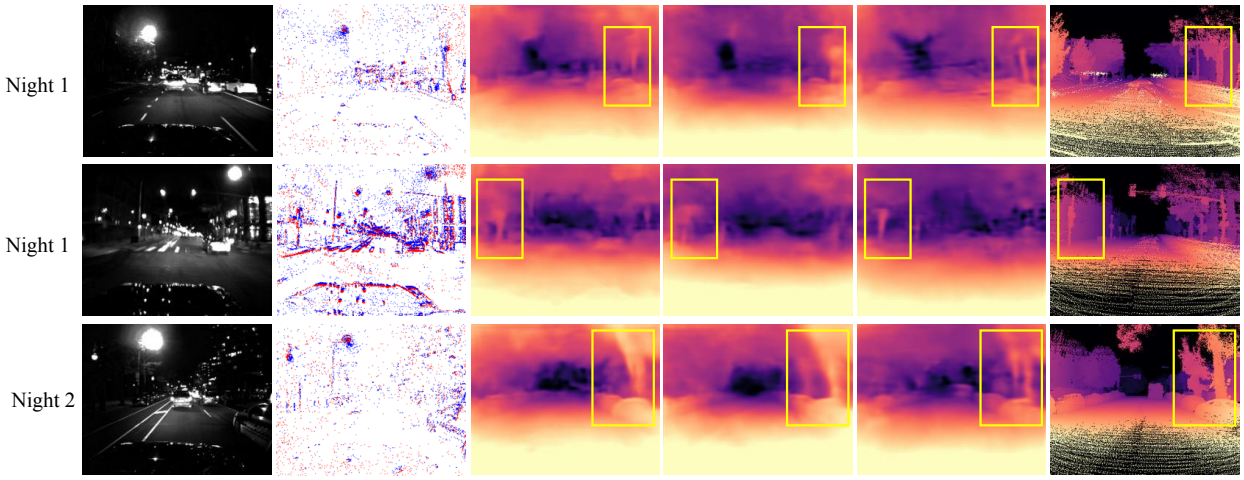
In Tab. II, we also provide a more detailed comparison with RAMNet, as it is the only open-source monocular depth estimation method that combines frames with events. Our SRFNet shows consistent superiority compared with RAMNet without pre-training in all input resolutions, sub-datasets and metrics. We also achieve competitive results with pre-trained RAMNet. For example, with 224×224 input resolution in the outdoor day1 sub-dataset, our method obtains 0.014 gain in the RMSE log metric. As shown in Fig. 4, our method predicts clearer trees in the first and third rows. Instead, RAMNet only predicts blurry structures on the trees. We ascribe it to the effectiveness of our proposed iterative feature fusion between frames and events.

We also conduct experiments on the synthetic DENSE



(a) Frame Input (b) Event Input (c) RAMNet (d) RAMNet [S] (e) Ours (f) Ground Truth

Fig. 4. **Qualitative comparisons with different methods for MVSEC dataset**. (c) RAMNet is purely trained on MVSEC; (d) RAMNet [S] denotes RAMNet pre-trained on synthetic dataset. The yellow bounding box indicates the region of significant contrast.



(a) Frame Input (b) Event Input (c) Baseline (d) w/ AIF (e) w/ AIF+RDR (f) Ground Truth

Fig. 5. **Qualitative results of ablation studies of SRFNet on the MVSEC dataset**. (c) denotes the baseline; (d) is our SRFNet without the RDR module; (e) is the complete SRFNet.

TABLE I

PERFORMANCE EVALUATION ACROSS MULTIPLE MAXIMUM CUT-OFF DEPTHS ON MVSEC USING AVERAGE ERROR (LOWER IS BETTER). NOTABLY, THE BEST RESULTS ARE HIGHLIGHTED IN BOLD, WHILE THE SECOND-BEST OUTCOMES ARE UNDERLINED. OUR METHOD EXCELS IN BOTH DAYTIME AND NIGHTTIME SCENARIOS AMONG METHODS EXCLUSIVELY TRAINED ON REAL-WORLD DATASETS. FURTHERMORE, [S] MEANS THE METHODS ARE PRE-TRAINED ON SYNTHETIC DATASETS.

		Frame-based		Event-based	Event and frame fusion			
Dataset	Distance	MegaDepth [13]	MonoDEVS [9]	E2Dpeth[S] [10]	RAMNet[S] [6]	RAMNet	Ours w/ AIF	Ours w/ AIF+RDR
outdoor day1	10m	2.38	1.47	1.02	1.05	1.29	0.97	**0.96**
	20m	3.63	2.49	1.93	1.99	2.23	**1.73**	1.77
	30m	4.54	3.13	2.47	2.63	2.89	**2.2**	2.37
outdoor night1	10m	4.64	2.99	1.78	2.31	6.31	1.47	**1.26**
	20m	7.55	3.71	2.45	2.92	6.03	2.08	**1.95**
	30m	8.8	5.08	3.26	3.64	6.71	3.04	**3.01**
outdoor night2	10m	4.28	1.77	**1.09**	1.11	1.09	1.26	1.19
	20m	7.04	3.17	2.20	2.26	2.93	**2.08**	2.13
	30m	8.39	4.66	3.36	3.49	4.26	**3.25**	3.22
outdoor night3	10m	3.71	1.40	**0.75**	0.88	0.91	1.07	1.01
	20m	6.50	3.01	**1.91**	2.31	2.35	1.98	2.12
	30m	7.79	4.68	**3.03**	3.79	3.95	3.34	3.52

dataset [10], following RAMNet [6], We set the $\alpha = 5.7$ and $D_{\max} = 1000$ in Eq. 4 for DENSE. As the quantitative comparison shown in Tab. IV, our SRFNet outperforms RAMNet under all of evaluation metrics. Additionally the visualization result comparison is depicted in Fig. 6, our SRFNet achieves more fine-grained prediction compared with RAMNet.

C. Ablation Study

AIF module. The proposed AIF module is based on the attention mechanism to interactively update the fused features and fused

TABLE II

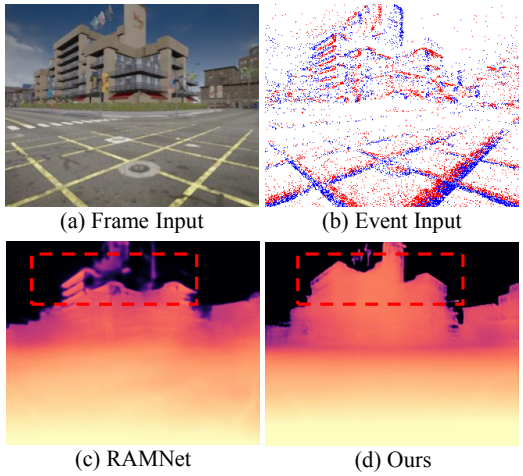
THE COMPARISON WITH PRE-TRAIN AND RETRAINED RAMNET UNDER MORE METRICS. ↓ DENOTES LOWER IS BETTER AND ↑ DENOTES HIGHER IS BETTER. EVEN WITHOUT THE HELP OF THE RDR MODULE, OUR METHOD STILL CONSISTENTLY OUTPERFORMS PRE-TRAINED RAMNET OR DEMONSTRATES COMPARABLE PERFORMANCE WITH INPUTS UNDER DIFFERENT RESOLUTIONS

Resolution	Dataset	Methods	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	SI log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
224×224	Outdoor day1	RAMNet [S]	0.246	0.162	**8.926**	0.378	**0.072**	0.593	0.806	0.899
		RAMNet	0.286	0.233	9.846	0.452	0.124	0.587	0.775	0.865
		Ours	**0.234**	**0.160**	9.150	**0.364**	0.073	**0.634**	**0.814**	**0.922**
	Outdoor night1	RAMNet [S]	0.431	0.739	**11.822**	**0.530**	**0.135**	**0.477**	0.664	**0.789**
		RAMNet	1.018	3.447	15.383	0.831	0.213	0.265	0.427	0.560
		Ours	**0.335**	**0.259**	13.608	0.544	0.143	0.465	**0.667**	0.787
256×320	Outdoor day1	RAMNet [S]	0.278	**0.207**	8.820	0.429	0.094	0.552	0.772	0.874
		RAMNet	0.335	0.364	9.498	0.507	0.158	0.552	0.747	0.842
		Ours	**0.268**	0.244	**8.453**	**0.375**	**0.079**	**0.637**	**0.810**	**0.900**
	Outdoor night1	RAMNet [S]	0.445	0.693	**10.408**	0.534	0.128	**0.434**	0.652	0.790
		RAMNet	1.011	3.490	13.222	0.808	0.236	0.291	0.466	0.597
		Ours	**0.371**	**0.316**	11.469	**0.521**	**0.120**	0.433	**0.662**	**0.800**

TABLE III

ABLATION STUDY CONDUCTED ON THE MVSEC DATASET TO EVALUATE THE DIFFERENT SETTINGS SRFNET BY USING AVERAGE ERROR. THE BEST RESULTS, HIGHLIGHTED IN BOLD, SHOWCASE THE EFFECTIVENESS OF EACH COMPONENT OF SRFNET ACROSS ALL ANALYZED METRICS

	AIF		RDR		day1			night1			night2			night3		
	Attention	Mask	SPN	Mask	10m	20m	30m	10m	20m	30m	10m	20m	30m	10m	20m	30m
Baseline					1.11	2.04	2.57	1.93	2.48	3.49	1.24	2.25	3.46	**1.00**	2.24	3.71
(1)	✓				1.07	1.85	2.29	2.11	2.51	3.58	1.72	2.44	3.58	1.53	2.34	3.59
(2)	✓	✓			0.97	**1.73**	**2.20**	1.3	1.97	3.07	1.26	**2.08**	**3.25**	1.07	**1.98**	**3.34**
(3)	✓	✓	✓		1.07	1.79	2.35	1.63	2.33	3.48	1.26	2.33	3.67	1.07	2.27	3.81
(4)	✓	✓	✓	✓	**0.96**	1.77	2.37	**1.26**	**1.95**	**3.01**	**1.19**	2.13	3.32	1.01	2.12	3.52



(a) Frame Input (b) Event Input

(c) RAMNet (d) Ours

Fig. 6. **Qualitative comparison with RAMNet on the DENSE dataset**; (c) is RAMNet trained on DENSE dataset; (d) is our SRFNet trained on DENSE dataset.

TABLE IV

COMPARISON WITH RAMNET ON SYNTHETIC DENSE DATASET [10].

Method	Avg. Error ↓				
	10m	20m	30m	Abs. Rel ↓	RMSE log ↓
RAMNet	2.619	11.264	19.113	1.189	0.832
Ours	**1.503**	**3.566**	**6.116**	**0.513**	**0.687**

masks. To verify its effectiveness, we first add the attention mechanism into AIF. The performance has an improvement in most scenes. However, in some scenes such as "night3", the performance drops as the direct fusion neglects the structural ambiguity between two modalities. Furthermore, as shown in Tab.III, by adding the spatial priors, *i.e.*, the mask learning, into the AIF module, the performance has an obvious improvement, *e.g.*, 0.12 gain in 20m metric on day1 sub-dataset. It demonstrates that learning the spatial reliability is powerful for the fusion of frames and events.

RDR module. To obtain the fine-grained depth estimation, we further add the RDR module. The RDR module is based on the SPN, whose confidence map is refined by the fused masks from the AIF module. In this case, we first simply employ the original SPN, and find that it less effective. We ascribe it that the direct predicting the relationships between pixels in the SPN is difficult, especially when the two modalities exhibit different spatial priors. Instead, by integrating the fused masks with the SPN, the performance can be improved such as in the night1 sub-dataset. As shown in Fig. 5, it can be found that by adding AIF module and RDR module, the trees in the three rows have clearer structures.

V. CONCLUSION

In this paper, we proposed a novel monocular depth estimation network, called SRFNet, which can effectively fuse the frames and events by considering the spatial reliability. This was based on the observation that frames and events might drop their reliability in the challenging conditions, such as high-speed motion scenes for the frames and scenes with marginal light change for the events. Therefore, direct fusing the two modalities would cause structural ambiguity and degrade the depth estimation performance. Accordingly, we proposed the AIF module to learn the consensus regions of the two modalities, which can guide the inter-modal feature fusion and enhance each modality's feature learning process. In addition, we proposed the RDR module to obtain fine-grained depth estimation by integrating the fused spatial priors from the AIF module. We demonstrated that our SRFNet outperforms existing SOTA methods in various scenes. Detailed ablation studies also show the superiority of the AIF module and RDR module, especially about the mask learning strategy.

Future work: We plan to generalize our fusion method to more tasks. We hope our thinking of the fusion technique can inspire the community for better combining the novel event sensors with traditional RGB cameras.

Acknowledgement: This paper is supported by the National Natural Science Foundation of China (NSF) under Grant No. NSFC22FYT45 and the Guangzhou City, University and Enterprise Joint Fund under Grant No.SL2022A03J01278

REFERENCES

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *ArXiv*, abs/1812.11941, 2018.

[2] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *AAAI Conference on Artificial Intelligence*, 2019.

[3] Hoonhee Cho and Kuk-Jin Yoon. Selection and cross similarity for event-image deep stereo. In *European Conference on Computer Vision*, 2022.

[4] Jonghyun Choi, Kuk-Jin Yoon, et al. Learning to super resolve intensity images from events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2768–2776, 2020.

[5] Huan Fu, Mingming Gong, Chaohui Wang, K. Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[6] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6:2822–2829, 2021.

[7] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3827–3837, 2018.

[8] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494, 2020.

[9] Akhil Gurram, Ahmet Faruk Tuna, Fengyi Shen, Onay Urfalioglu, and Antonio M. L'opez. Monocular depth estimation through virtual-world supervision and real-world sfm self-supervision. *IEEE Transactions on Intelligent Transportation Systems*, 23:12738–12751, 2021.

[10] Javier Hidalgo-Carri'o, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. *2020 International Conference on 3D Vision (3DV)*, pages 534–542, 2020.

[11] Sayed Mohammad Mostafavi Isfahani, Kuk-Jin Yoon, and Jonghyun Choi. Event-intensity stereo: Estimating depth by the best of both worlds. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4238–4247, 2021.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018.

[14] Zhi Li, Xiaoyang Zhu, Haitao Yu, Qi Zhang, and Yongshi Jiang. Edge-aware monocular dense depth estimation with morphology. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2935–2942, 2021.

[15] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *ArXiv*, abs/2203.04838, 2022.

[16] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. In *NIPS*, 2017.

[17] Xu Liu, Jianing Li, Xiaopeng Fan, and Yonghong Tian. Event-based monocular dense depth estimation with recurrent transformers. *ArXiv*, abs/2212.02791, 2022.

[18] Yunfan Lu, Zipeng Wang, Minjie Liu, Hongjian Wang, and Lin Wang. Learning spatial-temporal implicit neural representations for event-guided video super-resolution. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1557–1567, 2023.

[19] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *European Conference on Computer Vision*, 2020.

[20] Mengyang Pu, Yaping Huang, Qingji Guan, and Haibin Ling. Rindnet: Edge detection for discontinuity in reflectance, illumination, normal and depth. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6859–6868, 2021.

[21] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12159–12168, 2021.

[22] Peilun Shi, Jiachuan Peng, Jianing Qiu, Xinwei Ju, Frank P.-W. Lo, and Benny P. L. Lo. Even: An event-based framework for monocular depth estimation at adverse night conditions. *ArXiv*, abs/2302.03860, 2023.

[23] Mennatullah Siam, Sepehr Valipour, Martin Jägersand, and Nilanjan Ray. Convolutional gated recurrent networks for video segmentation. *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3090–3094, 2016.

[24] Irwin Sobel, Gary Feldman, et al. A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in*, pages 271–272, 1968.

[25] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European Conference on Computer Vision*, 2021.

[26] Lei Sun, Kailun Yang, Xinxin Hu, Weijian Hu, and Kaiwei Wang. Real-time fusion network for rgb-d semantic segmentation incorporating unexpected obstacle detection for road-driving images. *IEEE Robotics and Automation Letters*, 5:5558–5565, 2020.

[27] Abhishek Tomy, Anshul K. Paigwar, Khushdeep Singh Mann, Alessandro Renzaglia, and Christian Laugier. Fusing event-based and rgb camera for robust object detection in adverse conditions. *2022 International Conference on Robotics and Automation (ICRA)*, pages 933–939, 2022.

[28] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios. *IEEE Robotics and Automation Letters*, 3(2):994–1001, 2018.

[29] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3349–3364, 2019.

[30] Lin Wang, Yujeong Chae, and Kuk-Jin Yoon. Dual transfer learning for event-based end-task prediction via pluggable event to image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2135–2145, 2021.

[31] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *ArXiv*, abs/2108.05015, 2021.

[32] Dian xi Shi, Luoxi Jing, Ruihao Li, Zhe Liu, L. Wang, Huachi Xu, and Yi Zhang. Improved event-based dense depth estimation via optical flow compensation. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4902–4908, 2023.

[33] Zheyuan Xu, Hongche Yin, and Jian Yao. Deformable spatial propagation networks for depth completion. *2020 IEEE International Conference on Image Processing (ICIP)*, pages 913–917, 2020.

[34] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, 2018.

[35] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1725–1734, 2019.

[36] Jiyuan Zhang, Lulu Tang, Zhaofei Yu, Jiwen Lu, and Tiejun Huang. Spike transformer: Monocular depth estimation for spiking camera. In *European Conference on Computer Vision*, 2022.

[37] Jiaming Zhang, Kailun Yang, and Rainer Stiefelhagen. Issafe: Improving semantic segmentation in accidents by fusing event-based data. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1132–1139, 2020.

[38] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and B. Dong. Object tracking by jointly exploiting frame and event domain. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13023–13032, 2021.

[39] Xu Zheng, Yexin Liu, Yunfan Lu, Tongyan Hua, Tianbo Pan, Weiming Zhang, Dacheng Tao, and Lin Wang. Deep learning for event-based vision: A comprehensive survey and benchmarks. *arXiv preprint arXiv:2302.08890*, 2023.

[40] Zhuyun Zhou, Zongwei Wu, Rémi Boutteau, F. Yang, Cédric Demonceaux, and D. Ginhac. Rgb-event fusion for moving object detection in autonomous driving. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7808–7815, 2022.

[41] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay R. Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3:2032–2039, 2018.

[42] Yihao Zuo, Li Cui, Xin-Zhong Peng, Yanyu Xu, Shenghua Gao, Xia

Wang, and Laurent Kneip. Accurate depth estimation from a hybrid event-rgb stereo setup. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6833–6840, 2021.