

# Enhanced Event-based Dense Stereo via Cross-Sensor Knowledge Distillation

Haihao Zhang<sup>1,2</sup> Yunjian Zhang<sup>3</sup>† Jianing Li<sup>3</sup>† Lin Zhu<sup>2</sup> Meng Lv<sup>2</sup>  
 Yao Zhu<sup>3</sup> Yanwei Liu<sup>1</sup> Xiangyang Ji<sup>3</sup>

<sup>1</sup>Institute of Information Engineering, Chinese Academy of Sciences

<sup>2</sup>Beijing Institute of Technology <sup>3</sup>Tsinghua University

haihaozhang@bit.edu.cn sdtczyj@gmail.com lijianing@pku.edu.cn

## Abstract

Accurate stereo matching under fast motion and extreme lighting conditions is a challenge for many vision applications. Event cameras have the advantages of low latency and high dynamic range, thus providing a reliable solution to this challenge. However, since events are sparse, this makes it an ill-posed problem to obtain dense disparity using only events. In this work, we propose a novel framework for event-based dense stereo via cross-sensor knowledge distillation. Specifically, a multi-level intensity-to-event distillation strategy is designed to maximize the potential of long-range information, local texture details, and task-related knowledge of the intensity images. Simultaneously, to enforce the cross-view consistency, an intensity-event joint left-right consistency module is proposed. With our framework, extensive dense and structural information contained in intensity images is distilled to the event branch. Therefore, retaining only the events can predict dense disparities during inference, preserving the low latency characteristics of the events. Adequate experiments conducted on the MVSEC and DSEC datasets demonstrate that our method exhibits superior stereo matching performance than baselines, both quantitatively and qualitatively.

## 1. Introduction

Depth estimation based on binocular stereo vision is a classic problem in computer vision [15, 16, 20, 24, 29, 36, 44, 46, 59], which aims to determine the correspondence between points in a rectified image pair. Deep learning-based stereo matching methods perform well on various public benchmarks [2, 6, 9, 23, 24, 31–33, 52, 54–57]. However, due to the shortcomings of frame-based sensors in terms of frame rate and dynamic range, it is very difficult to apply traditional cameras in more challenging scenarios such as

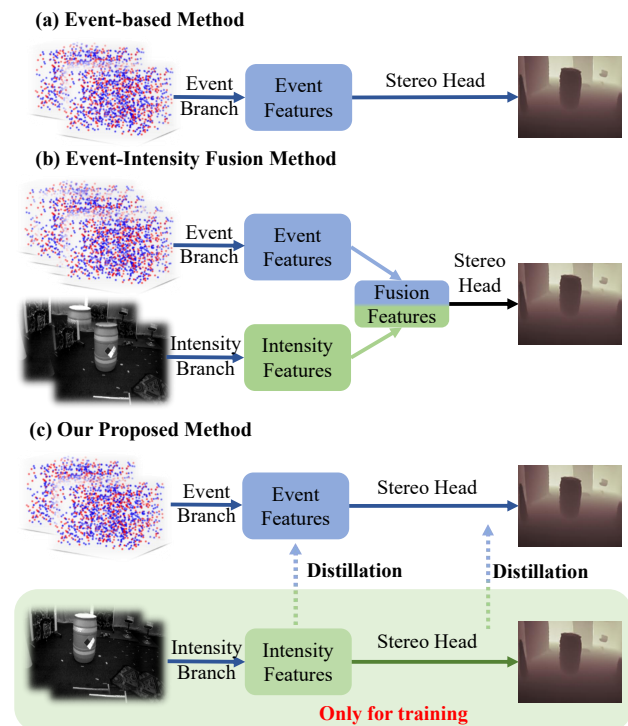


Figure 1. Comparisons on stereo matching frameworks (event-based, event-intensity fusion methods) and our proposed IED<sup>2</sup>S that presents an effective intensity-to-event distillation strategy. Texture and dense knowledge is transferred to the event branch, leading to dense stereo matching in the inference stage.

high speed and extreme lighting conditions.

Event cameras [3, 7, 30, 35, 40, 47, 64] are bio-inspired sensors and report intensity changes asynchronously in the form of an event stream, with each containing spatial location, polarity, and timestamps. They have the advantages of low latency and high dynamic range, making them immune to high-speed motion blur and can operate in extreme lighting scenarios. Therefore, they offer a satisfying solution to overcome the limitations of frame-based sensors.

Recent event-based stereo matching studies [1, 10, 11,

† Corresponding authors: Jianing Li, Yunjian Zhang

[13, 18, 37, 38, 49] embed asynchronous event data into image-like representations, then fine-grained features are extracted and subsequent disparity regression is performed. While disparity can be effectively estimated, since event cameras can only perceive changes in intensity, when the scene is stationary or there is little relative motion between the camera and the scene, the event stream will be very sparse. This makes it an ill-posed problem to obtain dense disparity maps using only sparse asynchronous event data.

Considering the dense features provided by intensity images, there raises a trend to combine events and intensity images for dense stereo matching [10, 11, 37]. They design algorithms that simultaneously input intensity images and events, using their complementary features to calculate disparities, and the results are outstanding advances for event-based stereo matching. Nevertheless, these approaches require the events and intensity images simultaneously during the inference stage, suffering from two problems: (1) The low latency characteristics of events is sacrificed unintentionally; (2) Processing both modalities concurrently consumes more computing resources, which is a heavy burden for real-time applications.

To address the above problems, we propose a novel framework named **Intensity-to-Event Distillation for Dense Stereo** matching method (**IED<sup>2</sup>S**), which **explores the potential of an efficient event-based stereo model for dense disparity prediction**. As shown in Fig. 1(c), we implement cross-sensor distillations from the intensity branch to the event branch. The dense and structural information of the intensity image provides useful regularization for asynchronous events, especially for ill-posed regions. In our framework, we design a novel feature alignment strategy, achieving **binocular-to-binocular** distillation for the first time. Specifically, a comprehensive multi-level intensity-to-event distillation strategy is designed, fully leveraging the modality-specific characteristics of shallow features and the task-related contextual understanding involved in deep abstract representations. Moreover, a warping-based intensity-event joint left-right consistency module is designed to ensure cross-view consistency. The intensity images are only used in the training stage. For inference, only event data is needed for obtaining dense disparity. Compared with fusion-based methods shown in Fig. 1(b), it has greater flexibility, preserving the low latency characteristics of event data and greatly reducing the inference burden. In summary, the main contributions of this work are:

- 1) We propose a dense event-based stereo matching method, **IED<sup>2</sup>S**, which is the first trial to implement binocular intensity-to-event distillation. The proposed **IED<sup>2</sup>S** enables the event branch to learn accurate dense information, greatly improving its disparity prediction capability.

- 2) We design a multi-level intensity-to-event distillation, fully distilling long-range information, local texture details,

and task-related knowledge into the event branch while avoiding the contamination of the original event modality.

- 3) Extensive experiments demonstrate that the proposed **IED<sup>2</sup>S** achieves significantly better performance than the prior works for event-based stereo matching.

## 2. Related Work

### 2.1. Event-Based Stereo

Early attempts used traditional hand-crafted methods [4, 12, 42–45, 68, 71] to determine the correspondence between two event streams. However, there are no fixed patterns in the event data, which makes it troublesome to use traditional methods to obtain dense disparity.

Recent learning-based approaches [1, 10, 11, 13, 37, 38, 49, 51] are able to estimate disparities using sparse events. The network usually includes feature extraction, cost aggregation, disparity regression, and disparity refinement modules. Tulyakov et al. [49] proposed a 4D queue embedding method that includes the spatiotemporal information of event data. Nam et al. [38] proposed an attention-based event concentration network to display scene details by omitting fewer details without covering events. DTC [62] used discrete time convolution modules to aggregate sequential event information at the feature level, accelerating the stereo matching process at a lower computational cost. Cho et al. [13] proposed the concept of stereo flow to aggregate past features and cost volumes. Accurate stereo matching is achieved by training the stereo matching and stereo flow simultaneously. However, it is an ill-posed problem to obtain dense disparity using only sparse asynchronous events. We therefore explore cross-sensor distillation for event-based dense stereo matching.

### 2.2. Cross-Sensor Distillation

Knowledge distillation is an effective method for compressing models while maintaining accuracy [21, 41, 50, 63]. In particular, cross-sensor knowledge distillation has received increasing attention and applied in many tasks, such as 3D object detection [8, 14, 22, 53, 65, 66], multispectral detection [34, 61], 3D hand pose estimation [60], and LiDAR semantic segmentation [25, 26, 48, 58]. Typically, a well-trained teacher model guides the student model’s feature extraction or result prediction that receives input from other modalities. Specifically, Hong et al. [22] first proposed cross-sensor knowledge distillation based on BEV, mimicking LiDAR BEV’s representation to achieve improved performance. Chong et al. [14] achieved the transfer of structural cues by projecting point clouds onto the plane of the image and maintaining similar local region affinities. Liu et al. [34] adopted different losses at multiple task levels to transfer coarse-grained and fine-grained features covering feature, detection, and segmentation. Tang et al. [48] estab-

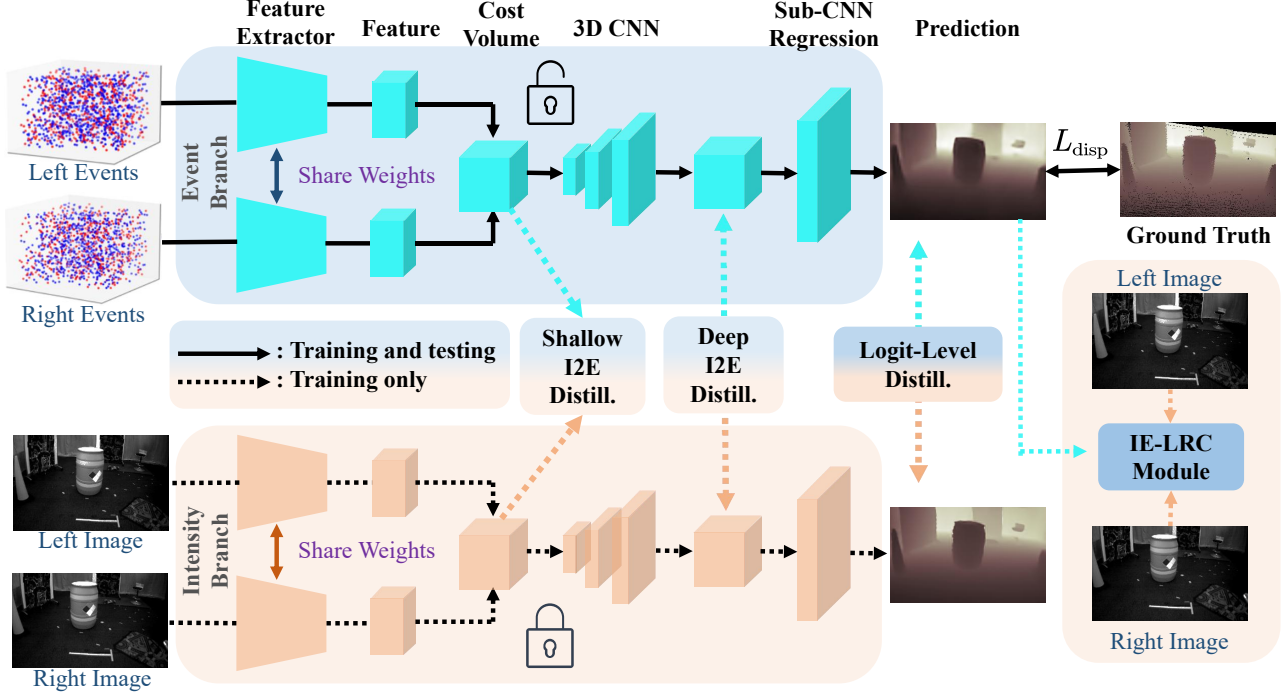


Figure 2. **The pipeline of the proposed Intensity-to-Event Distillation for Dense Stereo matching method (IED<sup>2</sup>S).** The proposed approach enables the event branch to learn a massive amount of texture knowledge and dense structure information under the guidance of the intensity branch. After training, the intensity branch is removed without introducing additional computational burden during inference.

lished a prototype-like library from strictly aligned fused features, and encouraged all point cloud features to learn from the prototypes.

Inspired by this, we propose an enhanced event-based dense stereo method via cross-sensor knowledge distillation. It obtains a massive amount of texture knowledge from the intensity images in the training stage, while in the inference stage, only events are needed to obtain dense and accurate disparity, greatly reducing the computational burden.

### 3. Methodology

#### 3.1. Problem Definition

The events of the left and right event cameras between  $t - 1$  and  $t$  are denoted as  $E_t^L$  and  $E_t^R$ , respectively, and the corresponding intensity images are indicated as  $I_t^L$  and  $I_t^R$ . The event data is a sequence of four-dimensional vectors  $(x, y, p, \tau)$ , where  $(x, y)$  means pixel coordinates,  $p \in \{-1, 1\}$  indicates polarity, and  $\tau \in (t - 1, t]$  represents continuous timestamps. We represent event data  $E_t^{L,R} \in \{(x, y, p, \tau) | t - 1 < \tau \leq t\}$  in voxel grid format following [70]. Specifically, we first scale the timestamps to bin indexes  $[0 : B - 1]$ , after which the events  $E_t^{L,R}$  are stacked to generate an voxel grid  $V_t^{L,R}(b, x, y) \in R^{B \times H \times W}$  with discretized time dimension  $B$ , where  $H$  and  $W$  represent the height and the width, respectively.

Given continuous stereo intensity image pair and stereo

event pair, with the same ground truth disparity, the training goal is to input four data in the training stage, which refer to  $V_t^L, V_t^R, I_t^L$ , and  $I_t^R$  respectively. By making full use of the information of the two modalities, in the inference phase, only events are needed to predict the dense disparity.

#### 3.2. Framework Overview

Fig. 2 shows a pipeline of our proposed framework. The framework consists of two identical branches, named the event branch and the intensity branch. The pre-trained intensity branch is used to distill knowledge to the event branch so that it can learn a massive amount of texture knowledge and dense structure information of intensity images. The distillation process includes three levels: shallow distillation, deep distillation, and logit-level distillation.

Shallow features contain much modality-specific information [53, 67], and distilling them into the event branch allows it to obtain complementary cross-sensor information. However, there are significant modality gaps between shallow features, such as textures of intensity images, and polarity and timestamps of events. The simple alignment would contaminate the original modality-specific information and introduce noise. To avoid this problem, we adopt a comprehensive distillation strategy from both long-range and focal perspective, thus the event branch can fully capture both long-range information and local texture details. For the deep features, we adopt a direct alignment strategy

because they mainly contain task-related contextual information. Furthermore, the logit-level distillation makes the disparity prediction of the event branch closer to the inspection prediction of the intensity branch. Besides the distillation strategy, to ensure cross-view consistency, we design an intensity-event joint left-right consistency module. This module attempts to enforce pixel-level consistency of the left intensity image with the warped one, which implicitly indicates the quality of the estimated disparity. In the inference stage, only keeping the event branch enables efficient and dense disparity prediction, and most importantly, preserving the low latency characteristics of events.

### 3.3. Multi-Level Intensity-to-Event Distillation

The proposed intensity-to-event distillation scheme enables the event branch to learn semantic cues from intensity images. It consists of three levels, including shallow distillation, deep distillation, and logit-level distillation. Shallow and deep distillation are both performed at the feature level, allowing the event branch to fully obtain long-range information, local texture details, and task-related knowledge. The last level drives the event branch to be close to the distribution of dense disparity of the intensity branch.

**Shallow Distillation.** In event-based stereo matching, sparse events lack detailed texture minutiae and rich semantic information, whether from the overall or local area details. As shown in Fig. 3, we extract **long-range** and **focal** knowledge from the event and intensity branch respectively, and then align them so that the event branch can obtain both long-range information and local texture details.

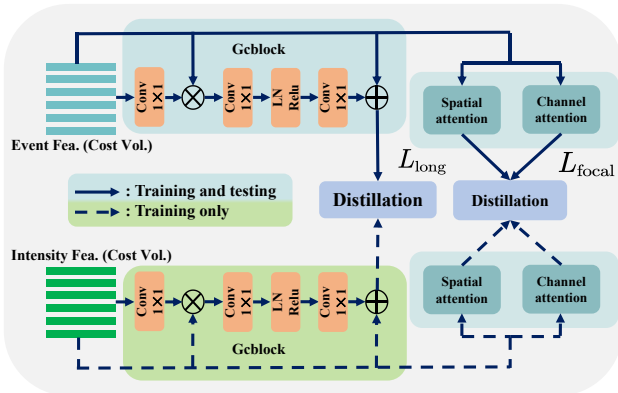


Figure 3. Illustration of shallow distillation. Features (cost volumes) of the two modalities are fed to generate relationship maps with long-range guidance. Simultaneous focal distillation is performed via channel attention and spatial attention. Shallow distillation is achieved through these two losses.

Long-range information distillation can transfer dark knowledge from the intensity branch to the event branch. However, there are significant modality gaps between shallow features, such as textures of intensity images, and polar-

ity and timestamps of events. To avoid the noise introduced by simple alignment, we use gcblock [5] to extract long-range relations ensuring that two branches have the same long-range relationships. The feature mapping with a long-range relation generated by gcblock is shown below:

$$M_{long}(x) = x \oplus (F(\ln(F(\sum_{i=1}^N \frac{e^{F(x_i)}}{\sum_{j=1}^N e^{F(x_j)}} x_i)))), \quad (1)$$

where  $F$  denotes the  $1 \times 1$  convolution layer,  $N$  means the number of pixels of the feature  $x$ ,  $\ln$  indicates the layer normalization, and  $\oplus$  denotes the broadcast element-wise addition. The long-range extraction loss can be obtained by computing the  $L_2$  norm between feature maps with long-range relationships as follows:

$$L_{long} = \|M_{long}(f_E) - M_{long}(f_I)\|_2, \quad (2)$$

where  $f_E$  and  $f_I$  represent the binocular features of the event branch and the intensity branch, respectively.

Long-range information distillation achieves the transfer of long-range information but still lacks local texture details. For stereo matching, focusing only on long-range relations is not sufficient to obtain dense disparity. We introduce spatial attention and channel attention to achieve intensity-to-event focal alignment. First, we compute the spatial and channel attention maps as follows:

$$A_s(x) = H \cdot W \cdot softmax(\frac{1}{C} \sum_{c=1}^C |f_c|), \quad (3)$$

$$A_c(x) = C \cdot softmax(\frac{1}{H \cdot W} \sum_{i=1}^H \sum_{j=1}^W |f_{i,j}|), \quad (4)$$

where  $A_s$  and  $A_c$  represent the spatial attention and channel attention maps derived from the features  $f$ . In order to enable the event branch to learn spatial and channel attention from the intensity branch comprehensively, we design the focal loss as follows:

$$L_{focal} = \beta_1 L1(A_s^E, A_s^I) + \beta_2 L1(A_c^E, A_c^I), \quad (5)$$

where  $\beta_1, \beta_2$  are hyper-parameters to balance the loss, and  $L1$  denotes the  $L_1$  norm.

**Deep Distillation.** As known from previous studies [48, 67], deep features are high-dimensional abstract representations with fewer modality variances. They retain rich task-related global contextual embeddings. In order to enable the event branch to effectively approximate the modality-independent intensity branch, we employ mean square error loss to facilitate deep distillation as follows:

$$L_{deep} = \sum_l^L \|F_E^l - F_I^l\|_2, \quad (6)$$

where  $l$  is the index where the distillation is performed.

**Logit-Level Distillation.** The prediction of the disparity regression layer represents the probability distribution of each modality branch. We propose the logit-level distillation, and then the disparity prediction of the event branch is learned from the soft labels generated by the intensity branch. We use Kullback-Leibler Divergence to measure the similarity in the probability distribution, making the disparity prediction of the event branch closer to the inspection prediction of the intensity branch, formulated as:

$$L_{logit} = KL(P_E, P_I). \quad (7)$$

In summary, the overall distillation loss  $L_{distill}$  is the sum of the long-range loss  $L_{long}$  (2) and focal loss  $L_{focal}$  (5) in the shallow distillation, the deep distillation loss  $L_{deep}$  (6), and the logit-level loss  $L_{logit}$  (7), as shown below:

$$L_{distill} = L_{long} + L_{focal} + L_{deep} + L_{logit}. \quad (8)$$

### 3.4. Intensity-Event Joint Left-Right Consistency

In order to implicitly indicate the quality of the estimated disparity and bring more explicit prompts, we introduce the **Intensity-Event joint Left-Right Consistency module (IE-LRC)**. For a stereo image pair, right image pixels can be shifted using disparity along each epipolar line to construct a warped left intensity image. Minimizing the difference between the warped left image and the real image reduces the ambiguity of the disparity [19]. The proposed IE-LRC module takes the disparity  $D^L \in R^{1 \times H \times W}$  predicted by the event branch, the left intensity image  $I^L \in R^{1 \times H \times W}$ , and the right intensity image  $I^R \in R^{1 \times H \times W}$  as input, as shown in Fig. 4. For stereo intensity image pairs, the right intensity image pixels are shifted using disparity predicted by the event branch along each epipolar to construct the warped left intensity image  $\hat{I}^L \in R^{1 \times H \times W}$ . The local appearance matching loss is constructed to minimize the reconstruction loss between the warped and real intensity image. We utilize both the structural similarity index loss and edge-aware smoothness loss to promote local smoothness of disparity, which can be expressed as follows:

$$L_{smooth} = \frac{1}{N} \sum_{i,j} |\partial_x D_{ij}^l| e^{-\|\partial_x I_{ij}^l\|} + |\partial_y D_{ij}^l| e^{-\|\partial_y I_{ij}^l\|}, \quad (9)$$

$$L_{lrc} = L_{smooth} + \alpha \frac{1}{N} \sum_{i,j} \frac{1 - SSIM(I_{ij}^l, \hat{I}_{ij}^l)}{2}. \quad (10)$$

We use a simplified SSIM with a  $3 \times 3$  block filter, and set  $\alpha = 0.85$ , following [19].

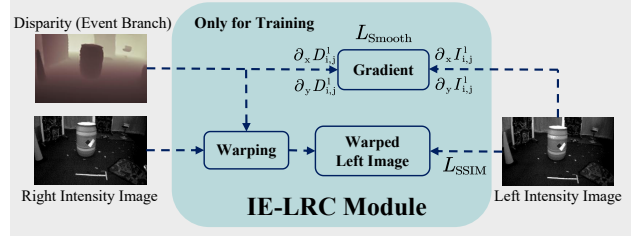


Figure 4. Illustration of intensity-event joint left-right consistency module. The input disparity comes from the event branch, and the input intensity images come from the intensity branch. The cross-view consistency across modalities is enforced through edge-aware smoothness loss and SSIM loss.

### 3.5. Training and Inference

**Overall Loss for Training.** The total training loss of our proposed framework consists of three components, namely the stereo matching loss  $L_{disp}$ , the intensity-to-event distillation loss  $L_{distill}$  (8), and the left-right consistency loss  $L_{lrc}$  (10), as shown below:

$$L_{total} = L_{disp} + L_{distill} + L_{lrc}. \quad (11)$$

The stereo matching loss  $L_{disp}$  can be defined as:

$$L_{disp}(D, D_E^*) = \frac{1}{V} \sum_{i=0}^V Smooth_{L_1}(d_i - d_i^*), \quad (12)$$

where  $D$  is the ground truth disparity,  $D_E^*$  is the disparity predicted by the event branch, and  $V$  is the number of valid pixels with ground truths,  $d_i$  for  $i \in \{1, 2, 3, \dots, V\}$  is the pixels in the ground truth disparity  $D$  that can be used for training, and  $d_i^*$  is the pixel at the corresponding position in the predicted disparity  $D_E^*$ . The smooth  $L_1$  loss function can be expressed as follows:

$$Smooth_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1, \\ |x| - 0.5, & \text{otherwise.} \end{cases} \quad (13)$$

**Inference.** The event branch is fully optimized during training, capturing a massive amount of texture knowledge and dense structure information from intensity images. During inference, retaining only the event branch can achieve efficient and dense disparity prediction without increasing the computational cost, and most importantly, preserving the low latency characteristics of events.

## 4. Experiment

### 4.1. Experimental Setting

**Datasets.** To evaluate the proposed method's effectiveness, the experiments are performed on two publicly available datasets, MVSEC [69] and DSEC [17], both from the real world. The MVSEC dataset has two DAVIS cameras

Method	Mean disp. error [pix] ↓		One-pixel accuracy [%] ↑		Mean depth error [cm] ↓		Median depth error [cm] ↓	
	Split1	Spilt3	Split1	Spilt3	Split1	Spilt3	Split1	Spilt3
EIS-E [37]	-	-	80.6	68.3	13.3	25.7	-	-
DDES [49]	0.59	0.94	89.8	74.8	16.7	27.8	6.8	14.7
DTC-PDS [62]	0.56	0.65	91.5	88.7	15.3	18.6	6.4	8.7
CTC-PDS [62]	0.53	0.73	91.6	88.2	14.9	20.6	6.4	10.6
EITNet [1]	0.55	0.75	92.1	89.6	14.2	19.4	5.9	10.4
DTC-SPADE [62]	<u>0.46</u>	0.60	<u>93.0</u>	89.7	13.5	17.1	5.2	7.9
TES [13]	<u>0.46</u>	<u>0.49</u>	92.9	<u>92.6</u>	<u>13.0</u>	<u>15.0</u>	<u>5.0</u>	<u>5.8</u>
Ours	<b>0.32</b>	<b>0.37</b>	<b>94.8</b>	<b>93.4</b>	<b>11.2</b>	<b>13.2</b>	<b>4.2</b>	<b>4.7</b>

Table 1. Disparity estimation results on the MVSEC [69] dataset. - indicates that results are not provided in the original paper. The best is in **bold** and the second best is in underlined.

with stereo setups that provide intensity images and pairing events. We use indoor flying from the MVSEC dataset, which was captured from a drone flying in a room with various objects. Following [10, 13, 49, 62], we split the MVSEC dataset indoor flying into three and use two of them, split 1 and split 3. For a fair comparison, we use mean disparity error, one-pixel accuracy, mean depth error, and median depth error as metrics. The DSEC dataset is a stereo event camera dataset for outdoor driving scenes. The intensity images and events are captured by different devices with different resolutions. All cameras are at the same height, so we approximately map the intensity images to the corresponding positions of the events with the provided camera matrix. For evaluation, we use the mean absolute error (MAE), one-pixel error (1PE), two-pixel error (2PE), and root mean square error (RMSE) as the metrics. See the supplementary material for more details on the dataset split.

**Implementation Details.** The proposed method is implemented using PyTorch and trained from scratch. We adopt the Adam [27] ( $\beta_1 = 0.9$ ,  $\beta_2 = 0.99$ ) optimizer to train the model with a batch size of 8 using  $256 \times 256$  random crops. We set the maximum disparity to 192. The initial learning rate is  $1 \times 10^{-3}$ , and the weight decay is  $1 \times 10^{-4}$ . All experiments are conducted on an NVIDIA RTX A6000 GPU. For more details about the implementation, please refer to the supplementary material.

## 4.2. Quantitative Results

We compare the results of the proposed **IED<sup>2</sup>S** with state-of-the-art methods. The comparison results are shown in Table 1. EIS-E [37], DDES [49], DTC-PDS [62], EITNet [1], and TES[13] only use asynchronous and sparse events, which limits the effectiveness of stereo matching due to the lack of sufficient texture knowledge and dense structure information. The proposed **IED<sup>2</sup>S** fully extracts extensive dense and structural information contained in intensity images and distills it into the event branch during the training, making its performance significantly better than existing event-based stereo matching methods. Specifically,

Methods	MAE ↓	1PE ↓	2PE ↓	RMSE ↓
DDES [49]	0.576	10.915	2.905	1.381
DTC-PDS [62]	0.526	<u>9.534</u>	<u>2.353</u>	1.263
Se-CFF [38]	<u>0.521</u>	9.586	2.623	<u>1.235</u>
Ours	<b>0.493</b>	<b>8.823</b>	<b>2.134</b>	<b>1.069</b>

Table 2. Disparity estimation results on the DSEC [17] dataset.

compared with the best-performing event-based method, TES[13], the proposed method outperforms it by 13.85% and 12.00% in terms of mean depth error for split1 and split3, respectively. We believe that this improvement comes from fully considering the modality gap and the novel design of the distillation strategy. Meanwhile, it is worth noting that the combined intensity images and events methods require both events and intensity images to be input at the inference stage, which limits the application of some scenarios, such as high-speed motion scenes or when intensity images are unavailable. However, ours only requires events as input during the inference phase, which fully utilizes the low-latency characteristics of events in special scenarios (high speed and extreme lighting).

Furthermore, we evaluate the proposed model on the DSEC dataset, and the results are shown in Table 2. Similarly, compared with the state-of-the-art baselines, ours achieves the best performance on all metrics and outperforms existing methods by a large margin, thanks to the designed cross-modal distillation in Sec. 3. Specifically, it outperforms Se-CFF [38] by 5.37% and 7.46% in terms of mean absolute error and one-pixel error, respectively, which verifies the effectiveness of the proposed method.

## 4.3. Qualitative Results

As shown in Fig. 5, we present a qualitative comparison of our method with the baseline in various scenarios on the MVSEC dataset. DDES [49] and TES [13] struggle to output dense disparity due to the lack of detailed information caused by the inherent sparsity of events. They are not ideal in some locations, such as the stripe lines dis-

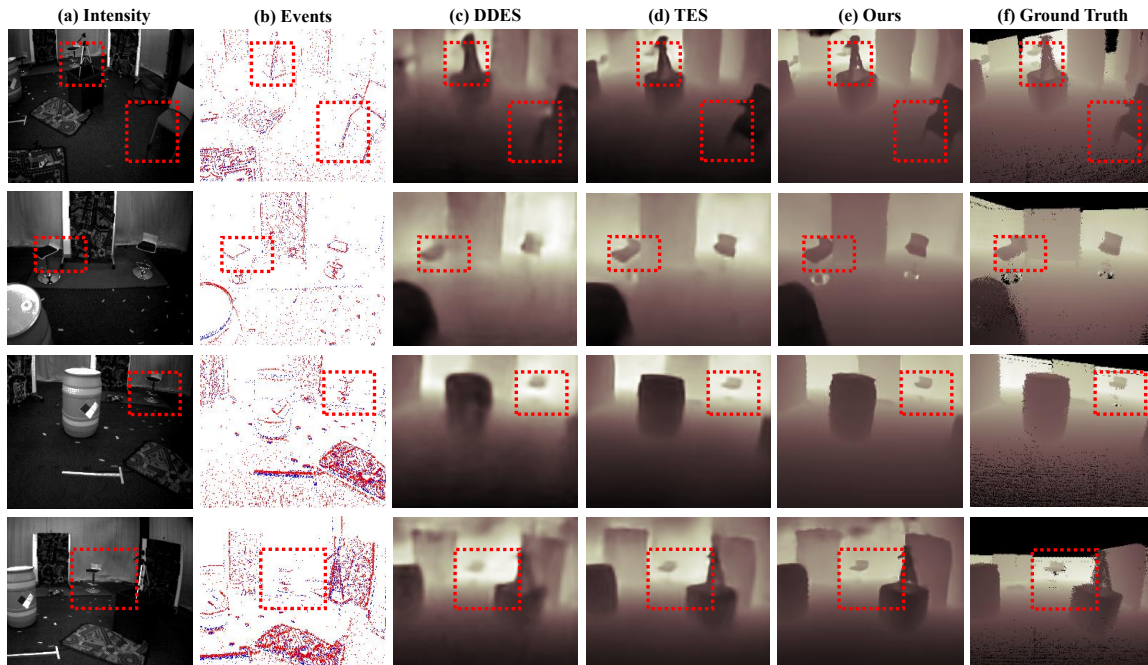


Figure 5. Qualitative comparison of dense disparity estimation for indoor flying scenes in the MVSEC dataset. The first two columns are the intensity images and the corresponding events for visual reference, where the red and blue discrete points represent positive and negative events, respectively. (c), (d), and (e) are the results of DDES [49], TES [13], and ours, respectively. Our proposed method fully uses the intensity image’s detail and texture information during the training stage, better displaying fine details in the region highlighted by the red box. Detailed disparity information is framed by the red box for comparison.

appearing where the disparity should change dramatically. For example, blur appears at the junction of objects such as barrels and chairs with the background. In contrast, our method addresses these challenges by enabling the model to capture more intricate features, thereby offering more reliable information for accurate disparity estimation in ill-posed regions. For example, for a chair placed far away, the proposed method achieves fine-grained disparity prediction. Consequently, even in areas where event data is sparse and closely resembles the surrounding environment, our approach produces disparities that are not only clearer and denser but also exhibit more precise boundaries, such as the area outlined by the red box in Fig. 5.

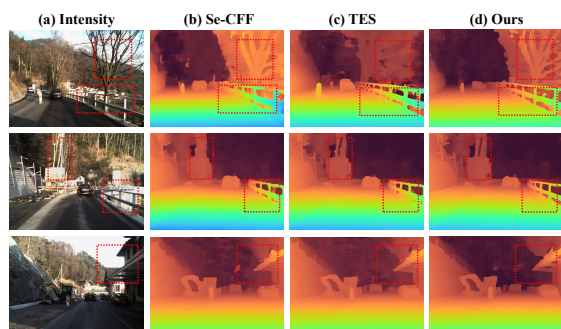


Figure 6. Qualitative results on the DSEC dataset. Our method can construct more dense disparities than others.

Additionally, we present results on the DSEC datasets in Fig. 6. We highlight regions with complex textures, where our method outperforms the others in predicting disparity with higher accuracy. For example, for objects such as trees and guardrails on the roadside, we obtain much sharper, more detailed, and artifact-free disparities.

#### 4.4. Ablation Studies

**Contribution of each component.** We perform extensive experiments on the MVSEC dataset to fully validate the effectiveness of each proposed component and related design choices in Sec. 3. We first present the results of the ablation study in Table 3. It can be observed that all the proposed components have contributed to the superior performance. This indicates that the information derived from intensity images is effective for dense event-based disparity estimation without significantly altering the network structure. When only one component is added, we see the most improvement comes from the feature distillation, which indicates that modality-specific and task-related features of the intensity branch are needed. When it is applied solely, the one-pixel accuracy is increased by 3.74% and 2.91% for split1 and split3, respectively. It is worth noting that intensity-event joint left-right consistency improves the one-pixel accuracy of split1 and split3 by 0.22% and 1.01%, respectively, which verifies the feasibility of improving the matching effect through ensuring cross-view consistency in

F-distill.	L-distill.	LRC	Mean disparity error [pix] ↓		One-pixel accuracy [%] ↑		Mean depth error [cm] ↓	
			Split 1	Split 3	Split 1	Split 3	Split1	Split3
✗	✗	✗	0.54	0.67	91.0	89.4	15.0	18.5
✓	✗	✗	0.38	0.46	94.4	92.0	12.1	13.9
✗	✓	✗	0.49	0.53	91.9	91.6	13.8	17.5
✗	✗	✓	0.50	0.64	91.2	90.3	14.6	18.0
✗	✓	✓	0.49	0.52	92.1	91.6	13.7	17.4
✓	✗	✓	0.37	0.41	94.5	92.7	11.6	13.7
✓	✓	✗	0.34	0.39	94.8	93.2	11.3	13.4
✓	✓	✓	0.32	0.37	94.8	93.4	11.2	13.2

Table 3. The result of our intensity-to-event distillation scheme with different settings on the MVSEC dataset. F-distill. denotes the feature distillation including shallow and deep distillation, L-distill. represents logit-level distillation, and LRC means left-right consistency.

stereo matching. When all components are combined, the performance is further improved without any conflicts, resulting in the best performance with a significant improvement of the one-pixel accuracy by 4.18% and 4.47% for split1 and split3, respectively.

Methods	Mean disparity error [pix] ↓		One-pixel accuracy [%] ↑	
	Split1	Split3	Split1	Split3
MSE	0.45	0.50	92.2	91.8
Si2eD	0.39	0.46	93.0	92.5
MSE+Di2eD	0.37	0.41	94.0	93.0
Si2eD+Di2eD	0.32	0.37	94.8	93.4

Table 4. The effectiveness of our designed feature distillation on the MVSEC dataset. Si2eD represents shallow distillation, and Di2eD denotes deep distillation.

**Feature alignment.** Although feature distillation has been a widely used loss term in many knowledge distillation studies [28, 39], we are the first to implement that for event-based stereo. For further exploration and discussion, we present our experiments to validate the design choice of the feature distillation module, as it brings the most improvement among all the proposed components. Specifically, to solely examine the effectiveness of our proposed shallow distillation, we compare it to the simple alignment feature-level distillation methods. MSE is a baseline method [22] that simply minimizes the element-wise  $L_2$  distance between the features of the teacher and student. As shown in Table 4, directly aligning the features of the two modalities contaminates the original modality-specific information and introduces noise to the event branch. Specifically, the one-pixel accuracy is reduced by 2.74% and 1.71% for split1 and split3, respectively. A possible reason for this performance gain comes from the large modality difference between events and intensity images, and the novel design fully exploits these modality-specific features while effectively avoiding cross-sensor contamination.

Setting	MVSEC → DSEC		DSEC → MVSEC	
Metrics	MAE ↓	IPE ↓	Mean disp. error [pix] ↓	One-pixel acc. [%] ↑
w/o distill.	0.782	12.413	0.74	79.8
Full model	0.563	9.973	0.46	91.4

Table 5. Cross-validation experimental results between two real-world datasets, verifying good generalization.

#### 4.5. Cross-Domain Generalization Performance

The modality-specific knowledge obtained by cross-sensor distillation is independent of the dataset and has good generalization under different domain distributions. In this section, we discuss the generalization performance of the proposed method on unseen domains. Specifically, we perform cross-validation on two real-world datasets, i.e., the model is trained on MVSEC (DSEC) and evaluated on DSEC (MVSEC). As shown in Table 5, the designed module is less affected by the domain gap. This is because the designed cross-modal distillation enables the event branch to acquire general knowledge for dense stereo matching.

## 5. Conclusion

We propose a novel framework for event-based dense stereo via cross-sensor knowledge distillation. We design a multi-level intensity-to-event distillation strategy, achieving binocular-to-binocular distillation for the first time. A massive amount of texture knowledge and dense structure information contained in the intensity image is distilled to the event branch through the designed strategy. At the same time, an intensity-event joint left-right consistency module is proposed to enforce cross-view consistency. Extensive experiments on the MVSEC and DSEC datasets demonstrate the effectiveness of the proposed framework.

## Acknowledgments

This work was supported in part by National Natural Science Foundation of China under grant No.62371450.

## References

- [1] Soikat Hasan Ahmed, Hae Woong Jang, SM Nadim Uddin, and Yong Ju Jung. Deep event stereo leveraged by event-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 882–890, 2021. 1, 2, 6
- [2] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *British Machine Vision Conference*, pages 1–11, 2011. 1
- [3] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A  $240 \times 180$  130 db 3  $\mu$ s latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 1
- [4] Luis A Camuñas-Mesa, Teresa Serrano-Gotarredona, Sio H Ieng, Ryad B Benosman, and Bernabe Linares-Barranco. On the use of orientation filters for 3d reconstruction in event-driven stereo vision. *Frontiers in neuroscience*, 8:48, 2014. 2
- [5] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 4
- [6] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 1
- [7] Shoushun Chen and Menghan Guo. Live demonstration: Celex-v: A 1m pixel multi-mode event-based sensor. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1682–1683. IEEE, 2019. 1
- [8] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qin-hong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*, 2022. 2
- [9] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in neural information processing systems*, 33:22158–22169, 2020. 1
- [10] Hoonhee Cho and Kuk-Jin Yoon. Event-image fusion stereo using cross-modality feature propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 454–462, 2022. 1, 2, 6
- [11] Hoonhee Cho and Kuk-Jin Yoon. Selection and cross similarity for event-image deep stereo. In *Proceedings of the European Conference on Computer Vision*, pages 470–486. Springer, 2022. 1, 2
- [12] Hoonhee Cho, Jaeseok Jeong, and Kuk-Jin Yoon. Eomvs: Event-based omnidirectional multi-view stereo. *IEEE Robotics and Automation Letters*, 6(4):6709–6716, 2021. 2
- [13] Hoonhee Cho, Jae-Young Kang, and Kuk-Jin Yoon. Temporal event stereo via joint learning with stereoscopic flow. In *Proceedings of the European Conference on Computer Vision*, pages 294–314. Springer, 2025. 2, 6, 7
- [14] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. *arXiv preprint arXiv:2201.10830*, 2022. 2
- [15] Weiqin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, David Suter, and Alireza Bab-Hadiashar. Semantic guided long range stereo depth estimation for safer autonomous vehicle applications. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):18916–18926, 2022. 1
- [16] Boyu Gao, Haoxiang Lang, and Jing Ren. Stereo visual slam for autonomous vehicles: A review. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 1316–1322. IEEE, 2020. 1
- [17] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 5, 6
- [18] Suman Ghosh and Guillermo Gallego. Event-based stereo depth estimation: A survey. *arXiv preprint arXiv:2409.17680*, 2024. 2
- [19] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 5
- [20] Rostam Affendi Hamzah, A Fauzan Kadmin, M Saad Hamid, S Fakhar A Ghani, and Haidi Ibrahim. Improvement of stereo matching algorithm for 3d surface reconstruction. *Signal Processing: Image Communication*, 65:165–172, 2018. 1
- [21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2
- [22] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *Proceedings of the European Conference on Computer Vision*, pages 87–104. Springer, 2022. 2, 8
- [23] Asmaa Hosni, Michael Bleyer, Margrit Gelautz, and Christoph Rhemann. Local stereo matching using geodesic support weights. In *16th IEEE International Conference on Image Processing*, pages 2093–2096. IEEE, 2009. 1
- [24] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE transactions on pattern analysis and machine intelligence*, 35(2):504–511, 2012. 1
- [25] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12605–12614, 2020. 2
- [26] Maximilian Jaritz, Tuan-Hung Vu, Raoul De Charette, Émilie Wirbel, and Patrick Pérez. Cross-modal learning for domain adaptation in 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1533–1544, 2022. 2
- [27] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

- [28] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. 8
- [29] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):1738–1764, 2020. 1
- [30] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Transactions on Image Processing*, 31:2975–2987, 2022. 1
- [31] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16263–16272, 2022. 1
- [32] Ximeng Li, Chen Zhang, Wanjuan Su, and Wenbing Tao. Iinet: Implicit intra-inter information fusion for real-time stereo matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3225–3233, 2024.
- [33] Biyang Liu, Huimin Yu, and Yangqi Long. Local similarity pattern and cost self-reassembling for deep stereo matching networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1647–1655, 2022. 1
- [34] Tianshan Liu, Kin-Man Lam, Rui Zhao, and Guoping Qiu. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):315–329, 2021. 2
- [35] Xu Liu, Jianing Li, Jinqiao Shi, Xiaopeng Fan, Yonghong Tian, and Debin Zhao. Event-based monocular depth estimation with recurrent transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1
- [36] Qian Long, Qiwei Xie, Seiichi Mita, Kazuhisa Ishimaru, and Noriaki Shirai. A real-time dense stereo matching method for critical environment sensing in autonomous driving. In *17th International IEEE Conference on Intelligent Transportation Systems*, pages 853–860. IEEE, 2014. 1
- [37] Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Event-intensity stereo: Estimating depth by the best of both worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4258–4267, 2021. 2, 6
- [38] Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Stereo depth from events cameras: Concentrate and focus on the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6114–6123, 2022. 2, 6
- [39] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5007–5016, 2019. 8
- [40] José Antonio Pérez-Carrasco, Bo Zhao, Carmen Serrano, Begona Acha, Teresa Serrano-Gotarredona, Shouchun Chen, and Bernabé Linares-Barranco. Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing—application to feedforward convnets. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2706–2719, 2013. 1
- [41] Mary Phuong and Christoph H Lampert. Distillation-based training for multi-exit architectures. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1355–1364, 2019. 2
- [42] Ewa Piatkowska, Ahmed Belbachir, and Margrit Gelautz. Asynchronous stereo vision for event-driven dynamic stereo sensor using an adaptive cooperative approach. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 45–50, 2013. 2
- [43] Ewa Piatkowska, Jurgen Kogler, Nabil Belbachir, and Margrit 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*, pages 53–60, 2017.
- [44] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time. *International Journal of Computer Vision*, 126(12):1394–1414, 2018. 1
- [45] Paul Rogister, Ryad Benosman, Sio-Hoi Ieng, Patrick Lichtsteiner, and Tobi Delbruck. Asynchronous event-based binocular stereo matching. *IEEE Transactions on Neural Networks and Learning Systems*, 23(2):347–353, 2011. 2
- [46] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47:7–42, 2002. 1
- [47] Waseem Shariff, Mehdi Sefidgar Dilmaghani, Paul KIELTY, Mohamed Moustafa, Joe Lemley, and Peter Corcoran. Event cameras in automotive sensing: A review. *IEEE Access*, 2024. 1
- [48] Pin Tang, Hai-Ming Xu, and Chao Ma. Prototransfer: Cross-modal prototype transfer for point cloud segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3337–3347, 2023. 2, 4
- [49] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1527–1537, 2019. 2, 6, 7
- [50] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1365–1374, 2019. 2
- [51] SM Nadim Uddin, Soikat Hasan Ahmed, and Yong Ju Jung. Unsupervised deep event stereo for depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7489–7504, 2022. 2
- [52] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024. 1

- [53] Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8637–8646, 2023. 2, 3
- [54] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Johann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023. 1
- [55] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12981–12990, 2022.
- [56] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023.
- [57] Gangwei Xu, Yun Wang, Junda Cheng, Jinhui Tang, and Xin Yang. Accurate and efficient stereo matching via attention concatenation volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1
- [58] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European conference on computer vision*, pages 677–695. Springer, 2022. 2
- [59] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2019. 1
- [60] Shanxin Yuan, Bjorn Stenger, and Tae-Kyun Kim. 3d hand pose estimation from rgb using privileged learning with depth data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2
- [61] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Low-cost multispectral scene analysis with modality distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 803–812, 2022. 2
- [62] Kaixuan Zhang, Kaiwei Che, Jianguo Zhang, Jie Cheng, Ziyang Zhang, Qinghai Guo, and Luziwei Leng. Discrete time convolution for fast event-based stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8676–8686, 2022. 2, 6
- [63] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3713–3722, 2019. 2
- [64] Bo Zhao, Ruoxi Ding, Shoushun Chen, Bernabe Linares-Barranco, and Huajin Tang. Feedforward categorization on aer motion events using cortex-like features in a spiking neural network. *IEEE transactions on neural networks and learning systems*, 26(9):1963–1978, 2014. 1
- [65] Haimei Zhao, Qiming Zhang, Shanshan Zhao, Zhe Chen, Jing Zhang, and Dacheng Tao. Simdistill: Simulated multi-modal distillation for bev 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7460–7468, 2024. 2
- [66] Lingjun Zhao, Jingyu Song, and Katherine A Skinner. Crkd: Enhanced camera-radar object detection with cross-modality knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15470–15480, 2024. 2
- [67] Runkai Zhao, Heng Wang, and Weidong Cai. Lanecmkt: Boosting monocular 3d lane detection with cross-modal knowledge transfer. In *ACM Multimedia*, 2024. 3, 4
- [68] Alex Zihao Zhu, Yibo Chen, and Kostas Daniilidis. Realtime time synchronized event-based stereo. In *Proceedings of the European Conference on Computer Vision*, pages 433–447, 2018. 2
- [69] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multi-vehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. 5, 6
- [70] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 3
- [71] Dongqing Zou, Feng Shi, Wei-Heng Liu, Jia Li, Qiang Wang, Paul KJ Park, and Hyunsurk Ryu. Robust dense depth maps generations from sparse dvs stereos. In *British Machine Vision Conference*, 2017. 2