Selection and Cross Similarity for Event-Image Deep Stereo

Hoonhee Cho and Kuk-Jin Yoon

Korea Advanced Institute of Science and Technology
{gnsgnsgml, kjyoon}@kaist.ac.kr

Abstract. Standard frame-based cameras have shortcomings of low dynamic range and motion blur in real applications. On the other hand, event cameras, which are bio-inspired sensors, asynchronously output the polarity values of pixel-level log intensity changes and report continuous stream data even under fast motion with a high dynamic range. Therefore, event cameras are effective in stereo depth estimation under challenging illumination conditions and/or fast motion. To estimate the disparity map with events, existing state-of-the-art event-based stereo models use the image together with past events that occurred up to the current image acquisition time. However, not all events equally contribute to the disparity estimation of the current frame since past events occur at different times under different movements with different disparity values. Therefore, events need to be carefully selected for accurate event-guided disparity estimation. In this paper, we aim to effectively deal with events that continuously occur with different disparity values in the scene depending on the camera's movement. To this end, we first propose the differentiable event selection network to select the most relevant events for current depth estimation. Furthermore, we effectively use feature-like events triggered around the boundary of objects, leading them to serve as ideal guides in disparity estimation. To this end, we propose a neighbor cross similarity feature (NCSF) that considers the similarity between different modalities. Finally, our experiments on various datasets demonstrate the superiority of our method to estimate the depth using images and event data together. Our project code is available at: https://github.com/Chohoonhee/SCSNet.

Keywords: Event cameras, stereo depth, multi-modal fusion

1 Introduction

Estimating depth from a stereo image pair has been an important problem in the field of computer vision [17,28]. Stereo-based depth estimation methods generally find correspondences for all pixels in the stereo pair image (*i.e.*, stereo matching) and estimate the depth through triangulation using camera parameters. It plays an important role in autonomous driving and augmented reality, which require 3D information of the scene.

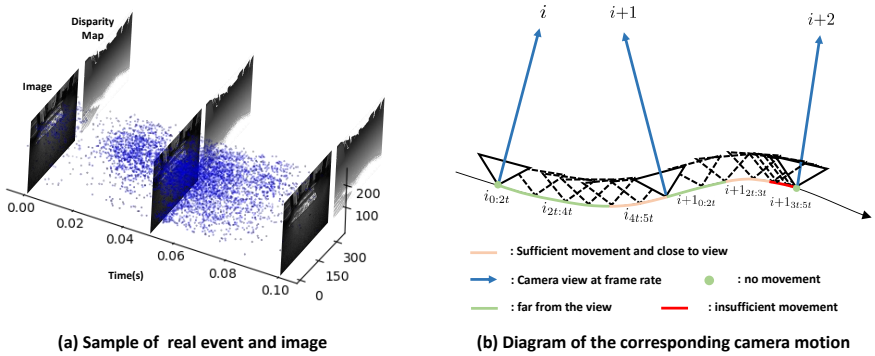

Fig. 1. Visualization of real-world events and images of indoor flying 1 in MVSEC datasets for stereo. For the disparity estimation of an image, the most recent events from the time that the image was acquired are used. (b) is a schematic diagram of the camera motion corresponding to (a). Here, i is the index of the frame, and t corresponds to 0.01 seconds in (a). Therefore, the time interval between the two frames corresponds to $5t$.

Most of recent stereo matching algorithms are learning-based and perform well on various large-scale public benchmarks [6,32,34,13,12,7]. However, there still exist some challenges in stereo matching because of the limitations of frame-based RGB sensors (*e.g.*, difficulty in operation in blurred or low dynamic range) and algorithmic incompletion [28] (*e.g.*, edge-fattening at depth discontinuities).

The event camera [4,19], a novel bio-inspired sensor, has provided a satisfactory solution to the limitations of frame-based RGB sensors in poor lighting or motion conditions. The event cameras asynchronously report the per-pixel changes of intensity in the form of a stream, called events. Event cameras have very low latency and cover a high dynamic range, making them intrinsically immune to motion blur and suitable for extreme lighting scenes. Therefore, the event camera can enrich the missing information as a complementary source for the shortcomings of the RGB sensor.

However, although event cameras can be a breakthrough to overcome the shortcoming of frame-based cameras, they require a significant transition in approach for real applications. Besides the new stereo matching algorithms for events, it is also essential to find optimal ways to represent the event data and put it into a depth estimation network, which has been a trend in event stereo research. In general, to estimate the depth map at a specific time using event cameras, the most recent events are used (see (a) of Fig. 1). Recently some stereo methods accumulate all events between two frames [29,2] or specify the number of recent events to stack [22]. However, these methods should make an assumption that the events corresponding to the same 3D scene point have the same disparity in the stack, which is not true in dynamic situations. In fact, the depth or disparity values of events can vary with time based on the movement of the camera and the scene. Therefore, depending on the motion of the camera

or the scene, events from different camera pose or object motion, which have inconsistent depth or disparity values, can be accumulated and used as input. To enable the network to extract relevant events for stereo in an end-to-end manner, we propose a differentiable event selection network. The differentiable event selection network extracts only relevant events among the all events accumulated between two frames with time information by using the image captured at a specific time as a condition to obtain accurate disparity. In addition, we regard the event data fired at the edge of the object or depth discontinuity as a feature, and propose the neighbor cross similarity feature that transfers information about the boundary of the object to the model from the event data refined through the selection network. We evaluate our method on the indoor event stereo dataset of MVSEC [39], and the outdoor public benchmark dataset of DSEC [10]. Furthermore, we present qualitative and quantitative comparisons for comparing with the state-of-the-art event stereo methods.

2 Related Works

2.1 Stereo depth estimation using images

The most successful methods of early studies using conventional RGB images have adopted end-to-end deep learning networks [32,37,18,35,34,12,6,33,30]. The networks generally comprise embedding, matching, and regularization modules. They outperform traditional methods by a large margin on public benchmarks (*e.g.*, Scene Flow [20] and KITTI [21]). However, the effects of motion blur and lightning on depth estimation remains a problem in terms of application.

2.2 Stereo depth estimation using events

With the rising of event cameras, attempts have been made to perform stereo depth estimation using temporally dense events. Using an event camera in the algorithm can solve the limitations of the RGB sensor, but finding a match between event pairs in the form of asynchronous streams is a remained issue. Early attempts utilized the hand-crafted method to determine corresponding events [15,5,41,42,24,27,38,25,26,8]. Their primary approach used window- or patch-based method defining the neighborhoods, and they succeeded in generating a depth map using spatial-temporal sparse event cameras. However, the characteristics of the event that do not follow the predefined pattern led to the inability to extract detailed and dense depth, and the performance was inferior to that of the learning-based method.

Recent methods improved the accuracy by adopting a learning-based approach capable of estimating dense depth with a sparse event. A new embedding of a 4D queue including both temporal and spatial information of event data for deep learning was proposed in [29], and the later study [2] improved the accuracy using various techniques. The state-of-the-art event stereo network [22] complemented spatially sparse events by using images together. They integrate

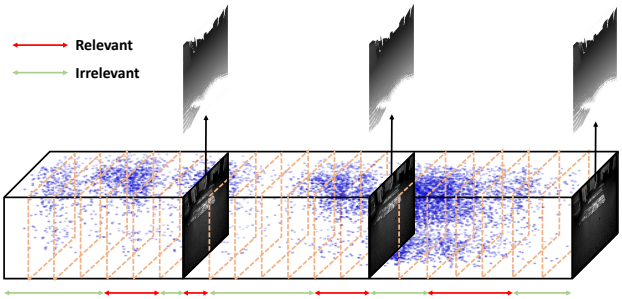

Fig. 2. Sample images and events from the MVSEC dataset [39]. The *relevant* event is generated by semantically related motion to the scene and is sufficient to estimate disparity. On the other hand, the *irrelevant* event corresponds to a view far from the scene of disparity to be obtained or contains insufficient information comprising real-world noise.

events and images through a recycling structure. Their method shows better performance than the setting that uses only an event or image data. However, they followed the existing event representation [31,9] used in the image reconstruction, and embedding for event-image disparity network has been less studied. Unlike high frame-rate image reconstruction, which requires only the most recent accumulated events for a short-time, in stereo matching, different threshold sensitivities between event camera pairs must be considered, and an event at an appropriate point in time according to motion and scene is required. In this paper, to tackle this problem, we explore a novel event embedding method for stereo depth estimation using events with images together.

3 Motivation

Existing event representation techniques that transform stream form of events to a machine-interpretable representation can be broadly classified into two categories: number- and time-based stacking. Number-based stacking methods [22,31,9] accumulate a certain number of events predefined by the user and put them into the network in the form of images. If a conservative amount of events is specified, insufficient events lead to ambiguity for matching, and conversely, too much amount gets rid of details and sharp edges. Also, even in the same dataset, the number of fired events varies depending on the scene. Because the number to be specified varies according to the distribution of the dataset, it is necessary to be able to access the dataset in advance, which may limit the application. The time-based stacking method [29,40,2] generally discretizes the event between two frames of an image along the time axis. Then, it converts it into a 3D or 4D image tensor for input to neural network architecture. The strength of this approach is that it is independent of the number of events triggered depending on the scene or motion. However, all events between two frames are not

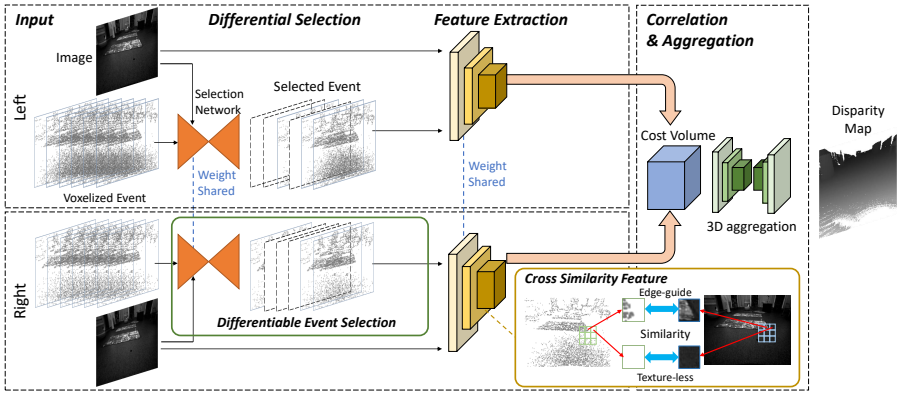

Fig. 3. The overview of architecture. In ***differential selection***, we select the most relevant event series from a voxelized event stack using a selection network (Sec. 4.1). We transform two inputs (events and images) of different modalities into a unified fused feature by using ***feature extraction*** that considers cross similarity features (Sec. 4.2). By utilizing ***correlation*** and ***aggregation*** (Sec. 4.3), we generate the final dense disparity output.

informative to obtain the disparity synchronized with the image. As depicted in Fig. 2., spatial–temporally sparse events cannot be preprocessed with a statically unified pattern.

The proposed event-image deep stereo algorithm is motivated by observations. First, all events between two frames are not required to estimate the time-synchronized depth with the image. Second, each pair of stereo event cameras has different threshold sensitivity. Even the stereo camera seeing the same view, the aspect of the event in each camera is other depending on the motion. Therefore, we propose a differentiable event selection network that selects relevant events to deal with a scene and motion variant issue.

4 Proposed Methods

Event Preparation We represent the stream format of events in the voxel grid format, considering both spatial and temporal coordinates. First, we collect events between two consecutive images. Then, following [40], we scale the timestamps to the range $[0, B-1]$ for inserting events into the discretized volume with the size of $w \times h \times B$ using a linearly weighted accumulation. Then, we can use the convolutional layer for the event data.

The overall framework of the proposed end-to-end depth estimation network is illustrated in Fig. 3. It consists of four sub-networks: differentiable event selection network, neighbor cross similarity feature extraction, correlation, and aggregation. Given a rectified image pair I_l, I_r and voxelized events E_l, E_r from stereo pairs of event cameras, using the differentiable event selection network,

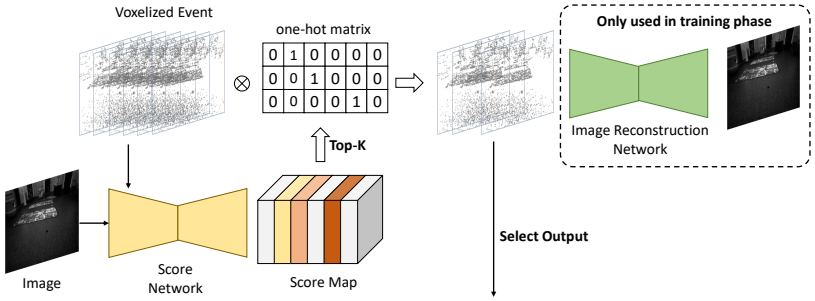

Fig. 4. Structures of the differentiable event selection network. Given a pair of images and events, the score network creates a score map to extract relevant events using images as conditions. Then, the selected event is inserted into the image reconstruction sub-network to restore the image and used for disparity estimation.

we extract the selected events S_l, S_r that are most relevant to the scene of disparity to be estimated. Next, the image and selected event are unified as a fused feature F_l, F_r through feature extraction. Then, 3D cost volume is constructed by correlating left fused feature F_l and right fused feature F_r. Finally, through the 3D aggregation network, we can obtain the dense disparity map.

4.1 Differentiable Event Selection Network

For the unity of denotation, we describe the left camera E_l, I_l as an example. The score network predicts a relevance score for each time region of the voxelized events concerning the image as the condition. The score map M is a grid with a size of $1\times1\times B$ as $M \in \mathbb{R}^B$, and each grid has a degree of relevance with the image. Given score map, we select the K most relevant event grids by creating a one-hot matrix $H \in \{0,1\}^{B\times \mathrm{K}}$. The one-hot matrix H consists of K number of one-hot vectors with a B dimension as $H = [h_1, h_2, \cdots, h_{\mathrm{K}}]$. Then, selected events can be extracted using a matrix multiplication as $S_l = E_l H$. Although these selected events can be directly used for depth estimation, we utilize an image reconstruction network to encourage the selection network to extract higher-quality events (see Fig. 4). The image reconstruction network is solely used during training to save computation time in inference without significantly increasing the memory footprint. The optimization of event selection network can be represented as a linear program of the form

$$\arg\max_{\boldsymbol{y}\in\mathcal{C}} \langle \boldsymbol{y}, M\mathbf{1}^\top \rangle, \tag{1}$$

where $M\mathbf{1}^\top \in \mathbb{R}^{B\times K}$ is score matrix obtained by multiplication, $\boldsymbol{y}$ is optimization variable, and $\mathcal{C}$ is a convex polytope set. However, one-hot operations from Top-K algorithms is non-differentiable, so we adopt the perturbed maximum method [3].

To define the gradients, add to input M a sampled random noise vector εZ, where $\varepsilon > 0$ is a hyper-parameter and Z has differentiable density $\mathrm{d}\mu(z) \propto \exp(-\nu(z))\mathrm{d}z$. Given the random perturbed inputs, expectations of results for each of n independent samples leads to smoothed versions as:

$$\boldsymbol{y}_\varepsilon = \mathbb{E}[\arg\max_{\boldsymbol{y}\in\mathcal{C}} \langle \boldsymbol{y}, M\mathbf{1}^\top + \varepsilon Z\rangle]. \tag{2}$$

In our experiment, we set $n = 200$ and $\varepsilon = 0.05$. From [1], for noise Z with $\mathrm{d}\mu(z) \propto \exp(-\nu(z))\mathrm{d}z$ and any twice differentiable ν, the Jacobian matrix of $\boldsymbol{y}_\varepsilon$ at M can be obtained as follows:

$$J_M \boldsymbol{y}_\varepsilon = \mathbb{E}[\arg\max_{\boldsymbol{y}\in\mathcal{C}} \langle \boldsymbol{y}, M\mathbf{1}^\top + \varepsilon Z\rangle \nabla_z \nu(Z)^\top / \varepsilon]. \tag{3}$$

Being able to compute the perturbed maximizer and its Jacobian allows optimizing functions that depend on M through $\boldsymbol{y}_\varepsilon$. Other distributions can be used in Z, but we use the Gumbel distribution [11] which is well-known in machine learning tasks. More details of the implementation for score network and image reconstruction network are provided in the supplementary material. This selection module applies equally to the right camera.

4.2 Neighbor Cross Similarity Feature Extraction

For the finding correspondences between left and right feature map, we should integrate the selected event S_l and image I_l into one feature F_l (For the right camera, this corresponds to S_r, I_r and F_r). Early study using events and images together for estimating depth in stereo setup [22] utilize the ***recycling network***, which recurrently stacks events and images. Their method shows better performance than the settings when using only events or images. However, because of the difference in modality between events and images, it is not practical to simply concatenate on the channel dimension or add in the entire feature dimension. Instead, we focus on the characteristic of events mainly fired at object edges as intensity changes usually happen. This characteristic can make the events ideal guidance to allow sharp depth values in boundaries, which is a challenging issue in image data. On the other hand, images have spatially dense characteristics, which can hand over the information about the entire scene to sparse and noisy events. According to this intuition, events and images can complement each other. To this end, we design feature extraction, including the Neighbor Cross Similarity Feature (NCSF) module, to extract representations that effectively include these correlations.

Our proposed feature extraction is demonstrated in details in Fig. 5. Feature extraction consists of cascaded Neighbor Cross Similarity Feature (NCSF) modules that contains the similarity between the two modalities. The NCSF module computes a similarity map using the cosine similarity between events and image features. Even if the image and the event are pixel-wise aligned, since the event changes temporally, unlike a static image, we consider the surrounding neighborhood together. Assuming that the size of the kernel considering

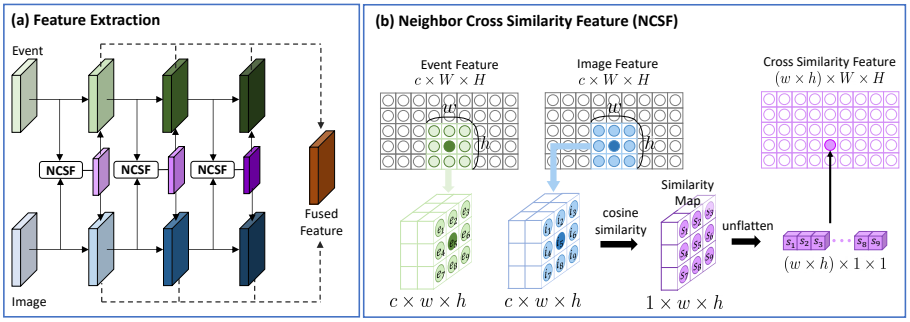

Fig. 5. The detailed structure of feature extraction. Feature extraction extracts the fused feature from an event and image pair from the identical camera. For correlating between two modalities, feature extraction consists of several cascaded Neighbor Cross Similarity Feature (NCSF) modules. NCSF modules generate cross similarity features by computing a similarity map between event and image features.

the neighborhood is specified as $w \times h$, the event feature in the kernel is represented as $\{e_1, e_2, \cdots, e_{w \times h}\}$ and the image feature corresponding to the pair is represented as $\{i_1, i_2, \cdots, i_{w \times h}\}$. Then, similarity map can be represented as $\{s_1, s_2, \cdots, s_{w \times h}\}$ and element of similarity map s_k can be obtained as follows:

$$\underset{k \in \{1, \cdots, w \times h\}}{s_k} = \left(\frac{e_k}{\|e_k\|_2} \right)^T \cdot \left(\frac{i_k}{\|i_k\|_2} \right) \tag{4}$$

Then, s_k is used as k-th channnel element in the cross similarity feature. In practice, we compute the cross similarity feature of every pair of pixels in the image and the event feature. To combine cross similarity features and event (or image) features, we design a fusion module for each modality consisting of a 1×1 convolution layer followed by BatchNorm and ReLU layers. In addition, we utilize this fusion module to extract the final fused features F_l. We concatenate all image and event features generated in the intermediate stage and apply the fusion module. In our experiments, we set the $w = h = 3$.

4.3 Correlation and Aggregation Network

For depth estimation, we need to correlate between the fused features pair F_l and F_r. By our proposed feature extraction, fused features contain much information, such as structure, boundary, and texture-less region. We adopt the group-wise correlation proposed in [12] to reduce information loss when obtaining the correlation between these two fused features. We set the number of groups N_g as 40. Besides, we utilize the stacked hourglass architecture proposed in [12]. They modified the stacked hourglass architecture proposed in [6]. There are four output modules for training, and only the last output module is adopted for inference. In each output module, the probability volume with a size of $D_{max} \times H \times W \times 1$

is generated using two 3D convolutions with upsampling and softmax function. The estimated disparity $\widetilde{\mathbf{D}}$ of each pixel can be obtained as follows:

$$\widetilde{\mathbf{D}} = \sum_{d=0}^{D_{max}-1} d \cdot p_d, \tag{5}$$

where d and p_d denote the possible disparity value and corresponding probability, respectively.

4.4 Objective Functions

We train our network in an end-to-end manner with three loss functions. Among them, two are for image reconstruction, and one is for disparity estimation.

Image reconstruction loss. In addition to the disparity estimation loss, we use the learned perceptual similarity loss (LPIPS) [36] and the L_1 loss as image reconstruction losses to train the differentiable event selection network more robustly. Through back-propagation in supervised learning with disparity ground-truth, although the selection network can be trained to some extent, we also utilize the image reconstruction, which has a much higher degree of relation to events than disparity estimation. As demonstrated in [36], the combination of these two losses encourage the sharp structural details. For LPIPS, we use variants of AlexNet [16]. We use the conv1-conv5 layers from [36].

Disparity estimation loss. We adopt the smooth L_1 loss function to train the proposed model. Smooth L_1 can be obtained as:

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{6}$$

The predicted disparity maps from the four output modules are denoted as $\widetilde{\mathbf{D}}_0$, $\widetilde{\mathbf{D}}_1$, $\widetilde{\mathbf{D}}_2$, $\widetilde{\mathbf{D}}_3$. Then,

$$\mathcal{L}_{disp} = \sum_{i=0}^{i=3} \Lambda_i \cdot \text{smooth}_{L_1}\left(\widetilde{\mathbf{D}}_i - \mathbf{D}^*\right), \tag{7}$$

where $\mathbf{D}^*$ denotes the ground-truth for the dense disparity map.

Our final loss ($\mathcal{L}$) is obtained by combining the image reconstruction losses ($\mathcal{L}_{L_1}, \mathcal{L}_{LPIPS}$) and the disparity estimation loss ($\mathcal{L}_{disp}$) as

$$\mathcal{L} = \mathcal{L}_{disp} + \lambda_1 \mathcal{L}_{L_1} + \lambda_2 \mathcal{L}_{LPIPS}. \tag{8}$$

5 Experiment

5.1 Datasets

We use two publicly available stereo real event camera datasets, the Multi-Vehicle Stereo Event Camera Dataset (MVSEC) [39] for indoor environments

and the stereo event camera dataset for driving scenarios (DSEC) [10] for outdoor environments.

MVSEC has a stereo setup with two DAVIS [4] cameras that can provide the images and pixel-wise aligned events with a resolution of 346×260 pixels. Following [22,29,2], we also use the *Indoor Flying* dataset from MVSEC, which is captured from a drone flying in a room with various objects, and partition them into three splits. We also do not evaluate the *split* 2 quantitatively due to the difference in dynamic characteristics in the training and testing events, as mentioned in [29,2]. For a fair comparison, we use the *mean depth error*, *mean disparity error* and *one-pixel-accuracy* used in [22,29,2] as the metrics.

DSEC provides high-resolution stereo event cameras captured in large-scale outdoor driving scenes. It contains 53 driving scenarios taken in various lighting conditions. However, DSEC does not provide an event aligned with the image; the event and the image are captured by different devices and have different resolutions and baselines. The image cameras with resolutions of 1440×1080 have a baseline of 4.5 cm with the event cameras with resolutions of 640×480, which leads to a disparity between pixels of events and images. Therefore, we approximately warp the image to the event location, which is not precisely aligned but can be used for evaluation. Since the ground-truth of the disparity map for the DSEC test set has not been publicly available, we evaluate the performance on the benchmark website. We use the *one-pixel error* (1PE), *two-pixel error* (2PE), *mean absolute error* (MAE), and *root-mean-square error* (RMSE) as the metrics.

5.2 Experimental Setup

We set the coefficients of Eq. 7 as $\Lambda_0 = 0.5$, $\Lambda_1 = 0.5$, $\Lambda_2 = 0.7$, and $\Lambda_3 = 1.0$, and the coefficients of Eq. 8 as $\lambda_1 = 1$ and $\lambda_2 = 1$. Our network is implemented with PyTorch [23]. We use the Adam [14] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. We set the voxelized event capacity B as 5 and the number of selections K as 3 in all experiments.

For MVSEC datasets, we train the our network for 30 epochs with a batch size of 2. The initial learning rate is set to 0.0001 without down-scale.

For DSEC datasets, we train the stereo networks for 120 epochs with a batch size of 8. The learning rate is set to 0.001 and down-scaled by 10 at 20, 40, and 60 epochs. The input voxelized event is randomly cropped with a size of 384×256 and vertical flip is applied for data augmentation. The input image is also cropped in proportion to the voxelized event size.

5.3 Quantitative Results

For quantitative analysis, we compare the results of our proposed model with the state-of-the-art method. There was no case of comparing with the frame-based methods on the MVSEC *indoor flying* dataset. Therefore, we train the frame-based model [12,6] using intensity images (APS) from the MVSEC dataset and select the models with the best performance in the validation until convergence.

Table 1. Results obtained for disparity estimation on MVSEC datasets. I indicates that the intensity image is adopted as the model modality, and E implies that the event data are adopted as the input. E + I means both conditions are adopted. The best and the second best scores are **highlighted** and underlined.

Model	Using Modality	Mean disparity error[pix] ↓		One-pixel accuracy [%] ↑		Mean depth error [cm] ↓	
		Split 1	Split 3	Split 1	Split 3	Split 1	Split 3
PSMNet [6]	I	0.57	0.68	88.6	83.1	15.9	18.3
GwcNet-gc [12]	I	0.53	0.64	89.9	85.8	15.0	17.4
PSN [29]	E	0.59	0.94	89.8	82.5	16.6	23.5
Ahmed *et al.* [2]	E	0.55	0.75	92.1	89.6	14.2	19.4
EIS [22]	E+I	–	–	89.0	88.1	13.7	22.4
Ours	E+I	**0.38**	**0.39**	**94.7**	**94.0**	**11.4**	**13.5**

Table 2. Results obtained for disparity estimation on DSEC datasets. (E) implies that the event data are adopted as the input, and (E+I) means both event and image are adopted. We report the results for each sequence in three different areas (Interlaken, Thun, and Zurich City) as well as the results of all sequences averages. The best and the second best scores are **highlighted** and underlined.

Model	1PE ↓				2PE ↓				MAE ↓				RMSE ↓			
	Inter	Thun	City	All	Inter	Thun	City	All	Inter	Thun	City	All	Inter	Thun	City	All
PSN (E) [29]	10.67	10.85	11.18	10.92	3.13	3.23	2.59	2.91	0.57	0.63	0.56	0.58	1.36	1.63	1.33	1.38
EIS (E+I) [22]	**4.77**	**5.15**	7.07	5.81	0.91	1.31	1.14	1.06	0.36	0.40	0.43	0.40	0.83	1.08	0.94	0.91
Ours (E+I)	4.86	5.30	**6.62**	**5.67**	**0.87**	**1.24**	**1.05**	**0.99**	0.36	0.40	**0.41**	**0.39**	**0.79**	**1.02**	**0.87**	**0.85**

Table 1 presents a comparison of the proposed method with previous single modality methods (only-events [29,2] or only-images [12,6]) and events-images fusion methods [22]. Our proposed modality fusion stereo network outperforms the earlier approaches by large margins in all evaluation metrics.

Furthermore, we also evaluate the proposed model on the DSEC dataset. Since DSEC's test dataset is not publicly accessible for the ground-truth disparity, we evaluate the network performance through the benchmark website. As presented in Table 2, we report the results for each sequence in different areas and the total average as well. Analytically, our method shows better performance on the overall sequence average than both the existing DSEC event-only baseline [29] and the state-of-the-arts [22] that use image and event fusion. Furthermore, especially in the Zurich City sequences that contain a lot of challenging illumination, such as Fig. 7, we show quantitatively better performance than the existing method through cross-modality similarity features between relevant events and images.

5.4 Qualitative Results

As shown in in Fig. 6, we qualitatively compare our results with previous works on MVSEC dataset. For comparison, we borrow the results of event-based approach [29] and event-image fusion method [22] from the original papers, respectively. In addition, we try our best to match the different color codings to those

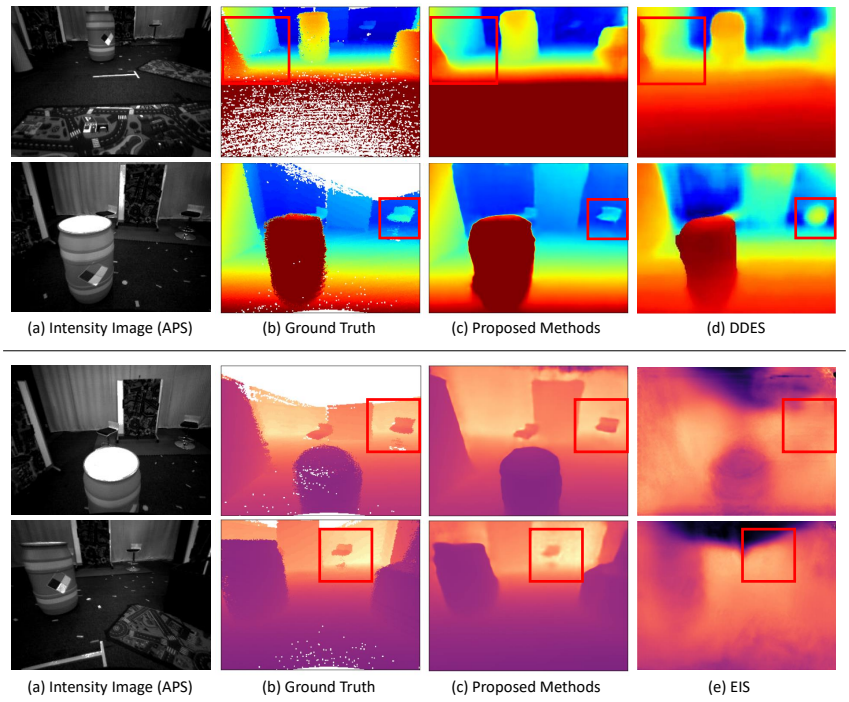

Fig. 6. Qualitative comparison of the proposed method with the previous methods from the MVSEC dataset. For comparison, we select frames similar to the ones used in [29] and [22]. Note that the results of DDES [29] and EIS [22] are borrowed from the original papers, respectively.

papers. As can be seen in the top row of Fig. 6 showing the comparison with the event-only method, our method using both modalities together allows for less noisy depth estimation. In addition, the bottom row shows the comparison with the event-image fusion method. Both our method and EIS use the same input of two types, different modalities. Still, our approach uses the event refined by the selection network and considers the correlation between modalities, estimating the artifact-free and much sharper results. In addition, we show the results of driving scene in challenging illumination conditions, which remains an open problem in stereo depth estimation, samples from the DSEC dataset. As can be seen in the highlighted region of Fig. 7, event data can capture an object with a high range covering an area that cannot be seen in the image. However, the event data also becomes noisier in situations such as night than in a general scene, and for this part, we supplement the context information from the image. Our strategy uses both event and image modalities to detect the depth of objects even under challenging illumination, which can be a breakthrough of direction that can solve the issues that remain in the conventional stereo matching from an application perspective.

Table 3. Ablation studies of the proposed components on the depth estimation.

Ablation Settings	Mean disparity error[pix] ↓		One-pixel accuracy [%] ↑		Mean depth error [cm] ↓	
	Split 1	Split 3	Split 1	Split 3	Split 1	Split 3
Baseline (GwcNet-g [12])	0.4020	0.4358	93.8809	91.7874	12.8391	16.5391
+ Differentiable Event Selection (DES)	0.4111	0.4153	94.1659	93.7863	12.3184	14.3000
+ Neighbor Cross Similarity Feature (NCSF)	0.3909	0.4095	94.1027	93.0166	11.6721	13.7153
Proposed (+ DES + NCSF)	**0.3776**	**0.3895**	**94.7201**	**94.0321**	**11.3645**	**13.4750**

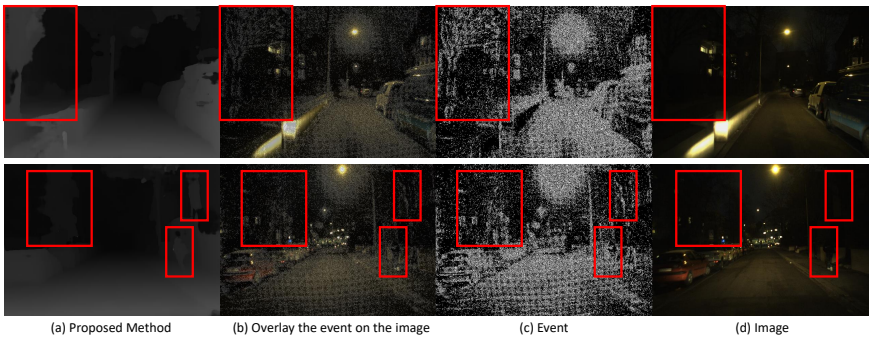

Fig. 7. Qualitative results of our proposed method from the challenging illumination scene on the DSEC dataset.

5.5 Ablation Studies

We perform ablation studies to confirm the effectiveness of the proposed methods using MVSEC dataset. Starting from the baseline, we add each sub-network to evaluate the performance. Since we adopt the 3D correlation network and aggregation network from [12], we use GwcNet-g as the baseline. In baseline, to combine the two modalities, event and image, we use a concatenation operation followed by a convolution layer. In Table 3, all of the proposed methods effectively improve the performance significantly.

5.6 Impact of the differentiable event selection

Results for the selection module are shown in Fig. 8. Unlike the method that fully accumulates between two frames, our method extracts the event most related to the boundary in the scene. Events refined by our differential event selection network, which is properly matched to object edges, can resolve discontinuous boundaries and be an ideal tool for estimating the sharp depth value. As mentioned in [29], in stereo depth estimation, spatial context is more reliable than temporal information, so our method that considers spatial correlation with images in continuous events leads to better depth results.

Furthermore, we analyze the effectiveness of the number of selection K and the voxel capacity B on the results of depth estimation in Table 4. When $B =$

Table 4. Impact of the voxelized event capacity B and the number of selection K on performance. The table shows the *one-pixel-accuracy* for *split* 1 test set in MVSEC dataset.

	The number of selection K				
voxel capacity B	1	2	3	4	5
5	91.6541	93.2581	**94.7201**	94.2207	94.1027
voxel capacity B	2	4	6	8	
10	92.6156	92.8351	<u>94.2313</u>	93.8791	

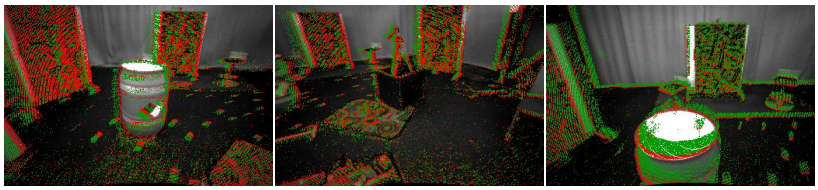

Fig. 8. The example of overlapping selected events and images. Selected events are shown in red. Except for the noise, most events are aligned to the object's boundary. Red: the selected events, green: the ignored events.

K = 5, it means using the entire voxel, and we analyze while reducing the K event select value. As the number of select decreases, the performance tends to improve, but when less than 3, the performance decreases. The reason for decreasing is that the amount of events is not sufficient to represent the overall scene, so it is challenging to obtain correspondence. In addition, when we increase the voxel capacity B to 10 and increase the number of selections K in proportion to B, the number of permutation cases increases significantly, and overall performance decreases. Still, it performs better than using the entire discritized voxel.

6 Conclusions

In this paper, we present the novel stereo depth estimation network using both modalities of events and images together. Specifically, we propose the differentiable event selection (DES) network to extract the events relevant to the scenes. Furthermore, we also propose a neighbor cross similarity feature (NCSF) that considers the similarity between different modalities. Finally, we evaluate our method with two real-world datasets, DSEC and MVSEC, and show the superiorness of our method in both quantitative and qualitative analyses. Our approach is effective for networks that use events and images together and can be generalized to other tasks.

Acknowledgements. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2022R1A2B5B03002636).

References

1. Abernethy, J., Lee, C., Tewari, A.: Perturbation techniques in online learning and optimization. Perturbations, Optimization, and Statistics p. 223 (2016)
2. Ahmed, S.H., Jang, H.W., Uddin, S.N., Jung, Y.J.: Deep event stereo leveraged by event-to-image translation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 882–890 (2021)
3. Berthet, Q., Blondel, M., Teboul, O., Cuturi, M., Vert, J.P., Bach, F.: Learning with differentiable pertubed optimizers. Advances in neural information processing systems **33**, 9508–9519 (2020)
4. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A 240 × 180 130 db 3 µs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits **49**, 2333–2341 (2014)
5. Camunas-Mesa, L.A., Serrano-Gotarredona, T., Ieng, S.H., Benosman, R.B., Linares-Barranco, B.: On the use of orientation filters for 3d reconstruction in event-driven stereo vision. Frontiers in neuroscience **8**, 48 (2014)
6. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 5410–5418 (2018)
7. Cheng, X., Zhong, Y., Harandi, M.T., Dai, Y., Chang, X., Drummond, T., Li, H., Ge, Z.: Hierarchical neural architecture search for deep stereo matching. ArXiv **abs/2010.13501** (2020)
8. Cho, H., Jeong, J., Yoon, K.J.: Eomvs: Event-based omnidirectional multi-view stereo. IEEE Robotics and Automation Letters **6**(4), 6709–6716 (2021). https://doi.org/10.1109/LRA.2021.3096161
9. Choi, J., Yoon, K.J., et al.: Learning to super resolve intensity images from events. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2768–2776 (2020)
10. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. IEEE Robotics and Automation Letters **6**(3), 4947–4954 (2021)
11. Gumbel, E.J.: Statistical theory of extreme values and some practical applications: a series of lectures, vol. 33. US Government Printing Office (1954)
12. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3268–3277 (2019)
13. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P.: End-to-end learning of geometry and context for deep stereo regression. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 66–75 (2017)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Kogler, J., Humenberger, M., Sulzbachner, C.: Event-based stereo matching approaches for frameless address event stereo data. In: International Symposium on Visual Computing. pp. 674–685. Springer (2011)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
17. Laga, H., Jospin, L.V., Boussaïd, F., Bennamoun: A survey on deep learning techniques for stereo-based depth estimation. IEEE transactions on pattern analysis and machine intelligence **PP** (2020)

18. Liang, Z., Feng, Y., Guo, Y., Liu, H., Chen, W., Qiao, L., Zhou, L., Zhang, J.: Learning for disparity estimation through feature constancy. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2811–2820 (2018)
19. Lichtsteiner, P., Posch, C., Delbrück, T.: A 128 × 128 120 db 15 µs latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-State Circuits **43**, 566–576 (2008)
20. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4040–4048 (2016)
21. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3061–3070 (2015)
22. Mostafavi, M., Yoon, K.J., Choi, J.: Event-intensity stereo: Estimating depth by the best of both worlds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4258–4267 (2021)
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
24. Piatkowska, E., Belbachir, A., Gelautz, M.: 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. pp. 45–50 (2013)
25. Piatkowska, E., Kogler, J., Belbachir, N., Gelautz, M.: 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. pp. 53–60 (2017)
26. Rebecq, H., Gallego, G., Mueggler, E., Scaramuzza, D.: Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time. International Journal of Computer Vision **126**, 1394–1414 (2017)
27. Rogister, P., Benosman, R., Ieng, S.H., Lichtsteiner, P., Delbruck, T.: Asynchronous event-based binocular stereo matching. IEEE Transactions on Neural Networks and Learning Systems **23**(2), 347–353 (2011)
28. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision **47**, 7–42 (2004)
29. Tulyakov, S., Fleuret, F., Kiefel, M., Gehler, P., Hirsch, M.: Learning an event sequence embedding for dense event-based deep stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1527–1537 (2019)
30. Tulyakov, S., Ivanov, A., Fleuret, F.: Practical deep stereo (pds): Toward applications-friendly deep stereo matching. arXiv preprint arXiv:1806.01677 (2018)
31. Wang, L., Ho, Y.S., Yoon, K.J., et al.: Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10081–10090 (2019)
32. Xu, H., Zhang, J.: Aanet: Adaptive aggregation network for efficient stereo matching. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1956–1965 (2020)

33. Yang, G., Zhao, H., Shi, J., Deng, Z., Jia, J.: Segstereo: Exploiting semantic information for disparity estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 636–651 (2018)
34. Zhang, F., Prisacariu, V.A., Yang, R., Torr, P.H.S.: Ga-net: Guided aggregation net for end-to-end stereo matching. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 185–194 (2019)
35. Zhang, F., Qi, X., Yang, R., Prisacariu, V., Wah, B., Torr, P.: Domain-invariant stereo matching networks. In: European Conference on Computer Vision. pp. 420–439. Springer (2020)
36. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
37. Zhong, Y., Dai, Y., Li, H.: Self-supervised learning for stereo matching with self-improving ability. arXiv preprint arXiv:1709.00930 (2017)
38. Zhu, A.Z., Chen, Y., Daniilidis, K.: Realtime time synchronized event-based stereo. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 433–447 (2018)
39. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters **3**(3), 2032–2039 (2018)
40. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–997 (2019)
41. Zou, D., Guo, P., Wang, Q., Wang, X., Shao, G., Shi, F., Li, J., Park, P.K.: Context-aware event-driven stereo matching. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 1076–1080. IEEE (2016)
42. Zou, D., Shi, F., Liu, W., Li, J., Wang, Q., Park, P.K., Shi, C.W., Roh, Y.J., Ryu, H.E.: Robust dense depth map estimation from sparse dvs stereos. In: British Mach. Vis. Conf.(BMVC). vol. 1 (2017)