*Journal of Imaging Science and Technology*® 66(6): 060402-1–060402-16, 2022.


# Self-Supervised Intensity-Event Stereo Matching

**Jinjin Gu**[†]
*School of Electrical and Information Engineering, The University of Sydney*

**Jinan Zhou**[†]
*Carnegie Mellon University*

**Ringo Sai Wo Chu**[†]**, Yan Chen, Jiawei Zhang, Xuanye Cheng, and Song Zhang**
*SenseTime Research and Tetras.AI*

**Jimmy S. Ren**
*SenseTime Research and Tetras.AI*
*Centre for Perceptual and Interactive Intelligence*
*E-mail: jimmy.sj.ren@gmail.com*

**Abstract.** *Event cameras are novel bio-inspired vision sensors that output pixel-level intensity changes in microsecond accuracy with high dynamic range and low power consumption. Despite these advantages, event cameras cannot be directly applied to computational imaging tasks due to the inability to obtain high-quality intensity and events simultaneously. This paper aims to connect a standalone event camera and a modern intensity camera so that applications can take advantage of both sensors. We establish this connection through a multi-modal stereo matching task. We first convert events to a reconstructed image and extend the existing stereo networks to this multi-modality condition. We propose a self-supervised method to train the multi-modal stereo network without using ground truth disparity data. The structure loss calculated on image gradients is used to enable self-supervised learning on such multi-modal data. Exploiting the internal stereo constraint between views with different modalities, we introduce general stereo loss functions, including disparity cross-consistency loss and internal disparity loss, leading to improved performance and robustness compared to existing approaches. Our experiments demonstrate the effectiveness of the proposed method, especially the proposed general stereo loss functions, on both synthetic and real datasets. Finally, we shed light on employing the aligned events and intensity images in downstream tasks, e.g., video interpolation application.* © *2022 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2022.66.6.060402]

## 1. INTRODUCTION

Event cameras measure changes in brightness at each pixel independently instead of reporting pixel activations. Event cameras have attracted increasing attention for their high temporal resolution, high dynamic range and low power consumption features, and have been applied to various computer vision tasks [1–5]. However, existing event cameras are either unable to obtain high-quality image pixel intensities (DVS [6] sensors only output events) or suffer low spatial resolution and lack of color information (dynamic and active-pixel vision sensors [7]). These limitations make it difficult for event cameras to assist computational imaging tasks, as we cannot obtain high-resolution intensity images and events simultaneously.

In this paper, we aim to connect a standalone event camera and a separate modern intensity camera so that applications could exploit the advantages of both sensors (see Figure 1). Such application scenarios are not uncommon for most consumer-level imaging devices, simply because acquiring colorful visual contents with high resolution, high speed, and low power consumption is without the scope of any individual image sensors. We establish a connection between these two sensors through a computational stereo matching model and estimate their disparity. This disparity describes the relationship between these two sensors and allows the sensors to be combined to complete the task that one sensor cannot achieve, e.g., obtaining both high-resolution images and events simultaneously for downstream tasks.

However, studying this problem is NOT a naive extension of the existing stereo matching methods on a new sensor setting, owing to the following technical barriers. First, the current stereo networks are not optimal for multi-modal problems. They assume that left and right view images have the same modality and use the weights shared feature extraction model for these images. Second, it is challenging to obtain multi-modal data for training as the acquisition of ground truth disparity for each different setting is expensive. To practically apply a multi-modal stereo framework, we need a robust training strategy and get rid of the shackles of data annotation. On this front, self-supervised learning provides a promising perspective to use the inherent

[†]Work was done when they were interns at SenseTime Research.

Received July 5, 2022; accepted for publication Oct. 28, 2022; published online Dec. 15, 2022. Associate Editor: Marius Pedersen.

1062-3701/2022/66(6)/060402/16/$25.00

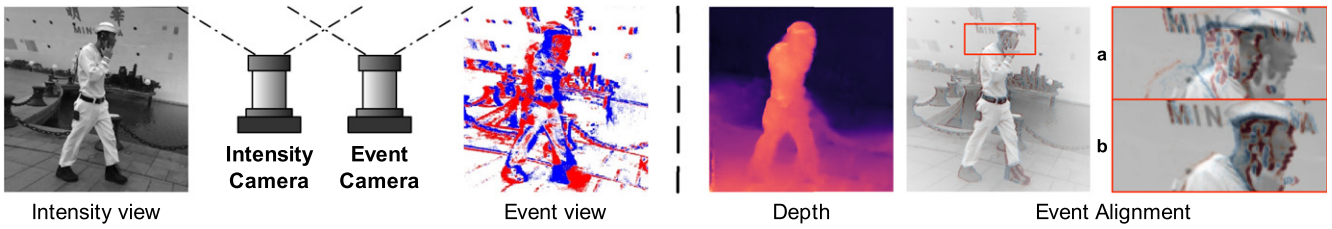


Figure 1. The proposed intensity-event stereo setting, in which we use an event camera and an intensity camera. With the proposed self-supervised stereo matching model, we can not only obtain the disparity used to calculate the depth map, but also build a connection between these two sensors. (a) The signals are from these two displaced sensors and are unaligned. (b) We align these two signals by the proposed method.

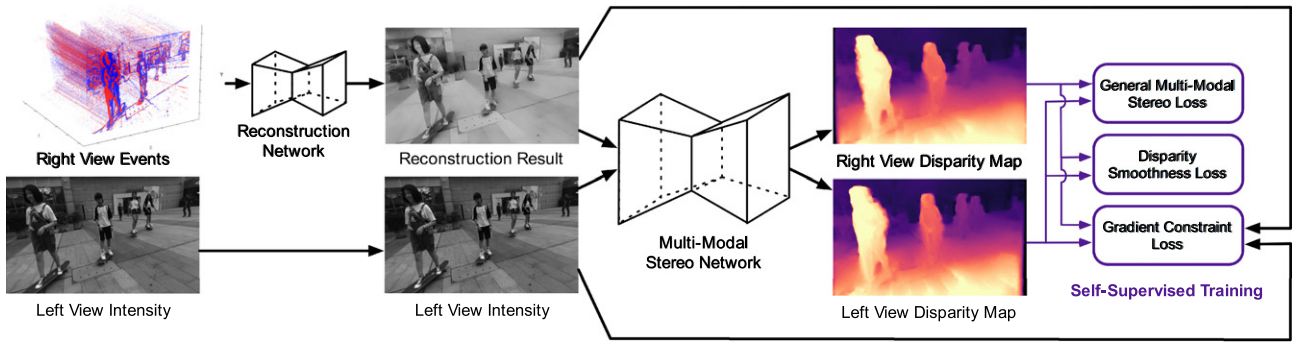


Figure 2. This figure shows the overall framework of the proposed method. We first obtain rough reconstruction results using the existing methods from the right view events. The multi-modal stereo network predicted a disparity map based on the right reconstructed results and left intensity images. The whole system uses three kinds of loss functions for self-supervised training.

constraints and characteristics of the data to learn the desired stereo matching without ground truth disparity indirectly. However, the prerequisite for the success of the existing self-supervised stereo matching framework is to establish the photometric consistency relationship between the projected images from two views. This brings up the third problem. The left and right view signals have different physical meanings and data structures in our setting. This causes the failure of the self-supervised learning framework as the previous photometric constraint does not hold.

In this work, we propose a self-supervised method for learning the multi-modal stereo matching without any ground truth disparity (see Figure 2). To facilitate the existing outstanding image stereo models on the proposed intensity-event setting, we first convert the event stream to roughly reconstructed images through the off-the-shelf models [8, 9]. The roughly reconstructed images are still in a different modality from the images of the other view as the color and detail information cannot be well restored. We improve the existing stereo networks and make images with different modalities to use modality-specific feature extraction sub-modules. In the proposed self-supervised method, we introduce a gradient structure consistency loss for the geometry constraints between the intensity and the reconstructed images after projection, which mainly leverages the edge information provided by events. Last but not least, only using the structure consistency may result in poor quality disparity maps as the supervision is sparse and vague. To overcome this issue, we propose a novel loss based on the cross-consistency between the disparities calculated across different views using different modality images. We also constrain our training according to the fact that the disparity of the same view should be zero. The proposed loss functions lead to improved stereo matching performance and robustness.

The calculated disparity maps can be used in many computational photography tasks, with depth estimation first. Projecting events to intensity camera view also allows many applications that could not be realized in the past due to hardware limitations. We can now obtain high-resolution events and intensity images simultaneously. At last, we experimentally demonstrate the potential of the proposed framework using the warped event to facilitate event-based video frame interpolation task.

## 2. RELATED WORK

### 2.1 *Event Cameras*

Event camera is a kind of sensor that records signals when the scene exhibits illumination changes [6, 7]. An event camera reports signals (events) asynchronously when the log intensity change exceeds a preset threshold $\tau$. We have witnessed the rise of event cameras due to their distinctive advantages over conventional active pixel cameras, e.g., higher frame rate, higher dynamic range and lower power consumption. These properties attracted the use of event cameras in many computer vision tasks, e.g., tracking

[1, 10], deblurring [2, 11], optical flow estimation [3, 12], SLAM [4, 13, 14], video frame interpolation [5, 11]. However, the unique data structure of event cameras renders the existing computer vision tools and algorithms unusable, which places a major obstacle against the application of event cameras. Many works have been focusing on bridging events and conventional cameras by reconstructing intensity frames from events [15–20], thus allowing modern vision algorithm to take place. Rebecq et al. [8] proposed E2VID, a recurrent network to reconstruct videos from a stream of events and trained it on a large amount of simulated event data. Scheerlinck et al. [9] proposed FireNet, which simplified the neural architecture in Ref. [8] with a smaller number of parameters while maintaining similar quantitative results. Although many studies have attempted to reconstruct the intensity image from the event, none of these methods can recover the intensity and color information well. Therefore, the absence of color information in the reconstructed image degrades the performance for downstream tasks. In our application, the color mismatch makes the existing self-supervised stereo matching algorithm based on photometric consistency invalid.

### 2.2 *Stereo Matching*

Stereo matching is the process of linking the pixels in different views that correspond to the same point of the scene. It follows a long line of research works. Early works involve searching and matching corresponding pixels on the epipolar line [21, 22]. Recently, deep learning based methods have dominated the field of stereo matching due to their superior performance and usability. Zbontar and LeCun [23] are among the first to use a convolutional network for computing stereo matching cost in image pairs. Following this, a number of studies were proposed to improve the performance, e.g., inner product layer [24], encoder-decoder architecture [25], 3D convolution cost-volume module [26], spatial pyramid pooling and 3D hourglass convolution [27], guided attention cost-column [28], PatchMatch module for sparse cost volume representation [29], intra-scaling cost aggregation [30]. With the development of various sensors, multi-model and cross-spectral stereo matching has become an emerging topic [31–36]. But none of them is suitable for calculating the correspondence between intensity images and events or event reconstruction images. Concurrent with our work, Mostafavi et al. [37] investigated stereo matching with event-intensity cameras on both views and proposed an event-intensity network that refines image details using events. Our work is essentially different in purpose and method; we use only one intensity and one event camera and train our model self-supervised.

### 2.3 *Self-Supervised Learning*

Learning-based stereo methods are data-hungry. They often require a lot of ground truth data for training. Over the past few years, self-supervised models have been developed to learn stereo matching without ground truth annotations. They are usually built on the principles of disparity smoothness prior and re-projection photometric consistency constraints. Garg et al. [38] tackled monocular depth estimation by minimizing the loss between the source image and backwards-warping from the subsidiary stereo image. Similarly, Godard et al. [39] included a left-right consistency to enforce disparity prediction. They further proposed a new minimum re-projection loss and auto-masking loss to improve the performance [40]. Zhou et al. [41] adopted left-right check to guide the training and pick suitable matching as training data. Zhi et al. [32] proposed a self-supervised learning framework for cross-spectral stereo matching. They introduced a material-aware loss function to handle regions with unreliable matching. However, their method involves the translation between intensity and near-infrared images and is thus unsuitable for our setting.

## 3. METHOD

In this section, we describe our self-supervised intensity-event stereo matching framework. We first introduce the problem formulation and overall framework design in Section 3.1. We then describe the modified stereo network for multi-modal problem in Section 3.2. The loss functions are introduced in Section 3.3, featuring a gradient structure consistency loss and general losses for multi-modal stereo matching.

### 3.1 *Overall Framework*

The multi-modal intensity-event stereo matching problem is first formulated as follows. As shown in Fig. 1, an intensity camera and an auxiliary horizontally displaced event camera are used in our setting. In this work, we assume that the camera on the left is the image camera, and the one on the right is the event camera. Let $I^l$ be the left view intensity image and $\{E^r_m\}_{m\in\mathbb{N}}$ be the event stream obtained by the right event camera within a short amount of time before the intensity image is captured. The underlying problem can be considered a data association problem, that is, to find correspondences between the points in the left image and right events. The correspondences are formed as the final disparity map, which is also the output of the stereo matching problem.

However, the right signal (event) is not in the same modality as the left signal (intensity images), failing the existing methods with the given problem setting. A reconstruction network is employed at first to convert the event stream $\{E^r_m\}_{m\in\mathbb{N}}$ to a roughly reconstructed image $I^r$. In our work, we use two popular event reconstruction models, E2VID [8, 42] and FireNet [9]. Note that these models are replaceable. Given these two images, we can adapt the existing image stereo models to predict disparity between images, which is also the disparity between the left image and the right events. However, the event camera does not record the value of the pixels and only records the pixels changing. Thus the rough reconstruction $I^r$ only contains usable edge information but are unreliable in color and detail and still in different modalities with $I^l$ (see the reconstruction result in Figure 3).

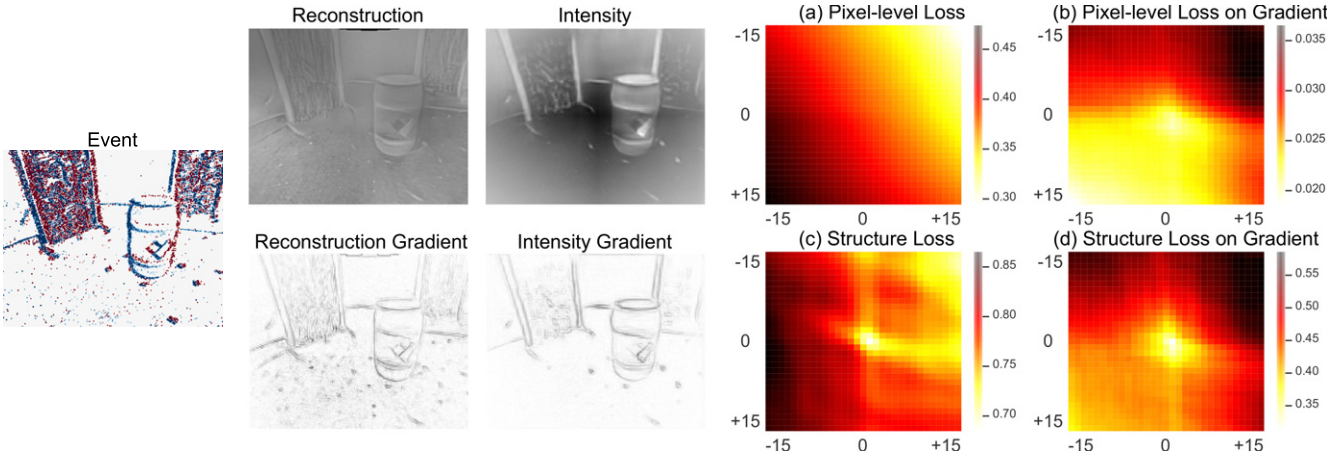


Figure 3. The visualization of different loss functions. On the left of the figure is the event, the reconstructed image using E2VID [8], the intensity image and their gradient visualization. The heat maps show the loss values with different displaced pixel numbers. In the centre of the heat maps, there is no displacement between the two images being compared, and the loss value should be the smallest.

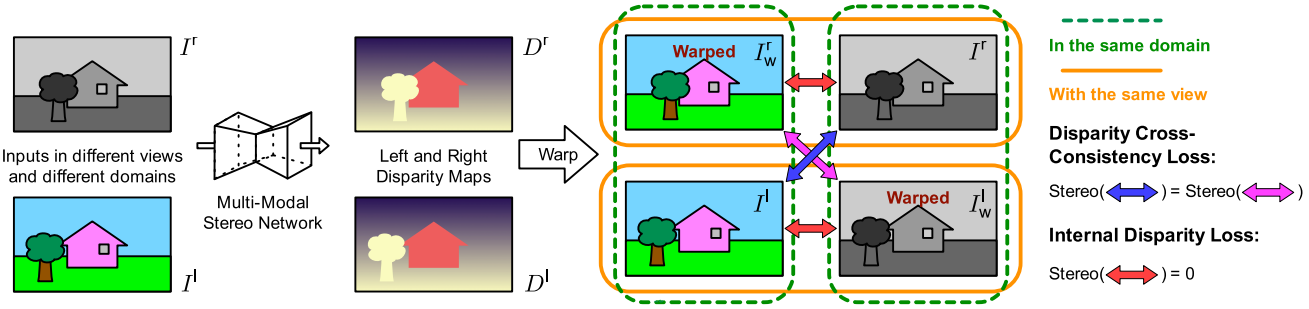


Figure 4. Using projection method and the multi-modal stereo model, we can obtain all four combinations of two views and two modalities. The proposed general multi-modal stereo losses are derived from the geometry constraints between these images, including disparity cross-consistency loss and internal disparity loss.

### 3.2 *Multi-Modal Stereo Network*

As stated above, we need a multi-modal stereo network to handle inputs with different modalities (either the right view is event voxel or reconstructed images). Although we can apply the previous convolutional stereo matching networks theoretically, the difference in modalities still poses challenges. Most stereo networks are composed of feature extraction, correlation and aggregation sub-models, and the feature extraction model usually share weights for both two views. This weight sharing strategy is effective originally but poses limitations for images with different modalities. We make the minor changes to these networks to make images with different modalities using modality-specific feature extraction sub-modules. This design has two advantages. First, the feature extraction models dedicated to different modalities avoid confusion between different images. Second, the new design allows us to swap the modalities of the left and right views and predict disparity of the other view, as long as we swap them together with the feature extraction branches. The second property is essential for the cross-consistency loss, which we will describe in Section 3.3. Note that this structure also allows the event voxel to be used directly as input.

### 3.3 *Loss Function*

To achieve the goal of self-supervised learning, we define the loss function as the combination of the following four parts:

$$\mathcal{L} = \lambda_{gd}\mathcal{L}_{gd} + \lambda_{sm}\mathcal{L}_{sm} + \lambda_{cc}\mathcal{L}_{cc} + \lambda_{itn}\mathcal{L}_{itn}, \qquad (1)$$

where the $\lambda$s denote the hyper-parameters that control the loss weights. Next, we describe each component of the loss function.

#### 3.3.1 *Gradient Structure Consistency Loss*

One core guarantee for the success of self-supervised stereo matching is that the output disparities indicate the epipolar geometry relationship between the left and right views. Usually, we can achieve this goal by comparing the projected left image and right images with image similarity metrics, e.g., pixel-wise loss and perceptual loss [43, 44]. However, the quality of the reconstruction result is usually poor, which fails the previous loss functions. Fig. 3 shows an example of reconstruction. As one can observe, the reconstruction network fails to recover any color information. However, its gradient reserves the structure information of the scene. We propose to use image structure loss [45] calculated on image gradient to constrain the stereo training. Let $G^l = \nabla_{xy} I^l$ be


Figure 5. Results of different methods on synthetic dataset. Monocular model does not work well as it cannot be fine-tuned under this setting. Stereo matching between intensity and reconstruction fails because the color discrepancy prevents the network to relate corresponding pixels. After self-supervised training, our predictions are on par with the ground truth. The reconstruction network is FireNet and the stereo network is the modified AANet.

the gradient of $I^l$ and $G^r = \rho \nabla_{xy} I^r$ be the adjusted gradient of $I^r$ with scaling factor $\rho$. The used gradient structure consistency loss is formulated as

$$\mathcal{L}_{gd} = 1 - \frac{2\mu_{G^l}\mu_{G^r} + c_1}{\mu_{G^l}^2 + \mu_{G^r}^2 + c_1} \times \frac{2\sigma_{G^l G^r} + c_2}{\sigma_{G^l}^2 + \sigma_{G^r}^2 + c_2}, \qquad (2)$$

where $c_1$, $c_2$ are constants and $\mu_{G^l}$, $\mu_{G^r}$, $\sigma_{G^l}$, $\sigma_{G^r}$, $\sigma_{G^l G^r}$ represent means, standard deviations and cross-covariance of the gradient pair. In practice, Eq. (2) is calculated on the local patch pairs and then summed up as the final loss.

We compare different losses in Fig. 3. We set the reconstructed image unmoved and shift the intensity images so that they are unaligned to simulate the situation after projection during self-supervised training. We then visualize the distribution of the loss values. We examine commonly used pixel-wise $l_1$-norm loss [39], pixel-wise $l_1$-norm loss on the image gradient, structure loss [39] on the image, and the proposed structure loss on the image gradient. As one can observe, the pixel-wise losses cannot indicate the optimal point. The structure loss calculated on images can indicate the optimal point but has an unsmooth loss landscape. Only the proposed loss has a relatively smooth loss landscape while successfully indicating the optimal point.

Directly calculating photometric consistency between warped images may introduce blurring to the predicted disparity map because there are occlusion areas between left and right view scenes, where warping cannot fill. We introduce the occlusion mask $M$ to mask out these occlusion pixels. We first perform left-right consistency check by projecting the right disparity using the left disparity map and calculate their coherence. The inconsistent region, which is likely to be the occlusion region, is marked as the occlusion mask $M$, which can be formulated as

$$M = \begin{cases} 0, & \|D^l - P(D^r; D^l)\|_1 < t \\ 1, & \|D^l - P(D^r; D^l)\|_1 \geq t, \end{cases} \qquad (3)$$

where $P(D^r; D^l)$ represents projecting the right disparity using the left disparity and $t$ is the threshold parameter.

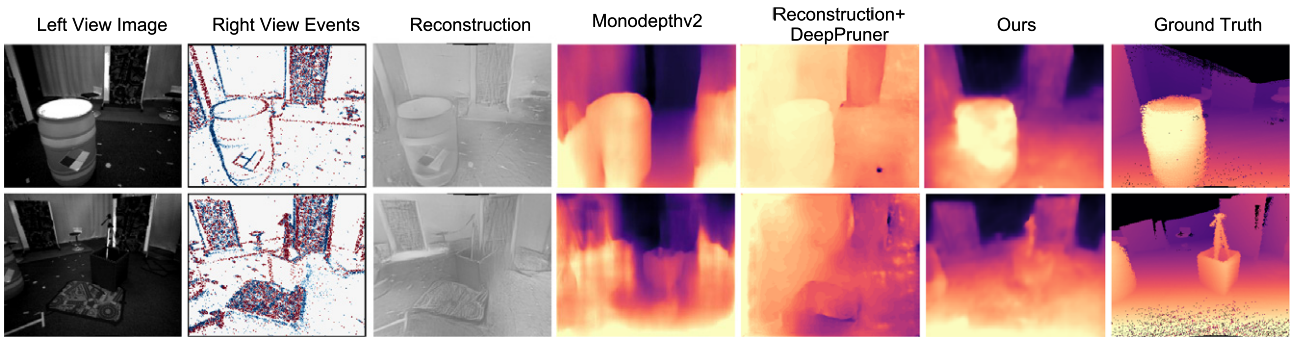


Figure 6. Comparison of different methods on the MVSEC dataset [49]. Notice that the direct event reconstruction quality on real data is far inferior to that of synthetics data, yet our framework can still achieve decent results, which shows its robustness. The reconstruction network is E2VID, and the stereo network is the modified AANet.

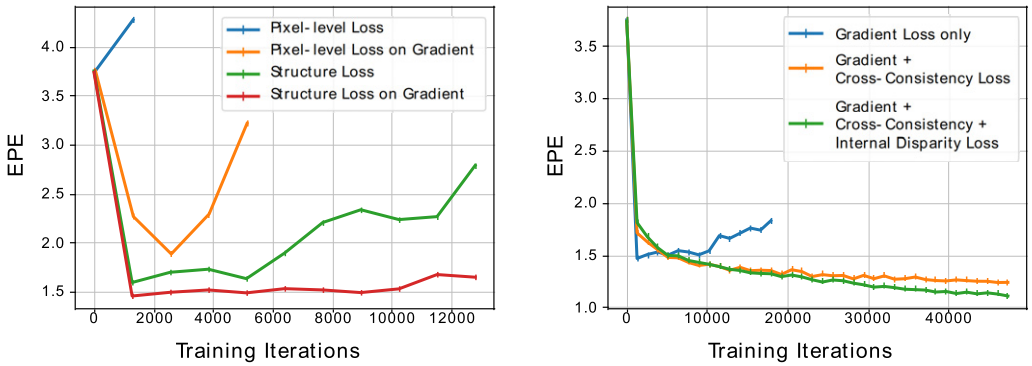


Figure 7. The convergence curves of using different losses. The left figure shows that the proposed gradient structure loss can provide meaningful self-supervision. The right figure indicates that the proposed general stereo loss could produce robust training effects. Some methods failed in the middle of training.

### 3.3.2 *Disparity Smoothness Loss*

After obtaining the disparity maps for both views $D^l$ and $D^r$, we follow the previous methods for estimating dense flow or disparity [39] and employ an edge-ware smoothness loss. This loss is symmetrical for both left and right views. Thus we omit the superscript. We encourage disparities to be locally smooth with a penalty on the disparity gradients $\nabla_x D$ and $\nabla_y D$. As depth discontinuities often occur at image edges, we weight this cost with an edge-aware term using the image gradients $\nabla_x I$ and $\nabla_y I$, which is formulated as

$$\mathcal{L}_{sm} = \frac{1}{N}\sum_{i,j} |\nabla_x D_{ij}| e^{-|\nabla_x I_{ij}|} + |\nabla_y D_{ij}| e^{-|\nabla_y I_{ij}|}, \qquad (4)$$

where $D$ denoted the disparity map corresponded with $I$, the subscripts $i$ and $j$ indicates pixel coordinates, $N$ is the total number of pixels.

### 3.3.3 *General Multi-Modal Stereo Losses*

Although gradient structure consistency loss can guide stereo training, the provided supervision is sparse and not that specific as pixel-level losses. We cannot obtain accurate disparity only with the above losses. Exploiting the internal stereo relationship between different views and different modalities, we propose general multi-modal stereo losses. An simple illustration of the proposed losses is shown in Figure 4. With the $I^l$ and $I^r$ at one hand, we calculate the disparities $D^l$ and $D^r$ that correspond to $I^l$ and $I^r$, respectively. By projecting $I^l$ and $I^r$ according to $D^l$ and $D^r$, we obtain $I^r_w$ and $I^l_w$, which represent different views and are in different modalities: $I^l_w$ is with the same modality with $I^r$ but with the same view with $I^l$; and $I^r_w$ is with the same modality with $I^l$ but with the same view with $I^r$. Using the same multi-modal stereo network, we can obtain $D^l_w$ and $D^r_w$ – the disparities calculated on two projected images. The proposed loss functions are built based on two facts. According to the fact that the disparity between $I^l_w$ and $I^r_w$ should be the same disparity between $I^l$ and $I^r$, we build the disparity cross-consistency loss to make them as close as possible:

$$\mathcal{L}_{cc} = \frac{1}{N}\sum_{i,j} \Big| |D^l| - |D^r_w| \Big| + \Big| |D^r| - |D^l_w| \Big|, \qquad (5)$$

where we take the absolute value for disparities as the projection directions may be opposite, and we only need their shapes. According to another fact that there should be no disparities within the same view, we build the internal

disparity loss:

$$\mathcal{L}_{itn} = \frac{1}{N}\sum_{i,j} |D^r_{itn}| + |D^l_{itn}|, \tag{6}$$

where $D^r_{itn}$ is the calculated between $I^r$ and $I^r_w$ and $D^l_{itn}$ is the calculated between $I^l$ and $I^l_w$.

## 4. EXPERIMENTS

### 4.1 *Implementation Details*

In this section, we experimentally evaluate the stereo matching performance of the proposed method. We use both synthetic and real data in our experiments. For the experiments based on synthetic data, we employ the Stereo Blur Dataset proposed by Zhou et al. [46], which contains 20,637 blurry-sharp stereo image pairs from 126 diverse sequences and their corresponding bidirectional disparities. To reliably synthesize events, we first increase the sequence frame rate from 60 fps to 2,400 fps via a high-quality frame interpolation algorithm [47] and then applying the V2E event simulator [48] to the high frame rate sequences. We use the officially split method, where 89 sequences are used for self-supervised training, and 37 sequences are used for testing. For real sensor data, we use the MVSEC [49] dataset, which contains the stereo intensity images and events captured by DAVIS 240C. MVSEC also provides ground truth depth captured by LiDAR. We use the official split method for the MVSEC dataset. Our method is implemented using Pytorch [50] framework and trained using NVIDIA V100 GPUs. For the stereo network design, we build our multi-modal networks by modifying DeepPruner [29] and AANet [30] according to Sec. 3.2. Note that our framework is compatible with the most alternative architectures of the reconstruction and the stereo networks. For optimization, we use Adam [51] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and learning rate $1 \times 10^{-4}$. We set the weighting of the different loss components to $\lambda_{gd} = 1$, $\lambda_{sm} = 0.1$, $\lambda_{cc} = 0.025$ and $\lambda_{itn} = 0.005$. The settings of $\rho$ and $t$ are experimental. The final value of $\rho$ is 1 on the synthetic dataset and 1.5 on the real world data. The value of $t$ is 2 for all datasets. The overall self-supervised training costs about 2 days.

### 4.2 *Stereo Matching Results*

We first quantitatively demonstrate the effectiveness of the proposed method in Tables I and II. The metrics are averaged end-point error (EPE) and >1-pixel, >3-pixel and >5-pixel error. Under the proposed setting, only limited methods can be used to obtain disparity maps as there is no multi-modal stereo matching model for event-intensity setting. We included the monocular depth model [40] for comparison, which was not fine-tuned on the target data, as we cannot obtain intensity images from both left and right views at the same time in this setting theoretically. We also consider the monocular event depth model [52], but the advantage of the event model is to combat high-speed motion and low-light situations that the intensity camera cannot handle, and its effect cannot be compared with the results predicted using the intensity images. We can have the following observations. Firstly, the proposed method can achieve much better results compared with monocular disparity estimation results using monodepthv2 [40] that is only based on the left image. It indicates that the information of another view plays an essential role in stereo matching. The unsatisfactory effect prevents us from using this disparity map to align events with the intensity image. We then show the results of directly performing stereo matching using the reconstructed right image and left image (marked as the "baseline experiments"). As can be seen, since the reconstructed images have a huge color difference against the left view

**Table I.** Quantitative comparison of different approaches on stereo matching using our synthetic stereo event dataset [46]. ↑ means the higher the better while ↓ means the lower the better. "*" indicates using the modified stereo network.

| Model | EPE ↓ | Bad Pixels ↓ | | |
|---|---|---|---|---|
| | | $\delta > 1$ | $\delta > 3$ | $\delta > 5$ |
| Monodepth2 | 8.849 | 0.953 | 0.781 | 0.648 |
| DeepPruner (upper bound) | 0.712 | 0.123 | 0.027 | 0.015 |
| FireNet+AANet (baseline) | 4.811 | 0.649 | 0.419 | 0.336 |
| E2VID+AANet (baseline) | 5.154 | 0.673 | 0.440 | 0.379 |
| FireNet+DeepPruner (baseline) | 10.29 | 0.417 | 0.226 | 0.181 |
| E2VID+DeepPruner (baseline) | 6.386 | 0.381 | 0.184 | 0.140 |
| FireNet+AANet* ($\mathcal{L}_{gd}$ and $\mathcal{L}_{sm}$) | 1.591 | 0.366 | 0.139 | 0.088 |
| E2VID+AANet* ($\mathcal{L}_{gd}$ and $\mathcal{L}_{sm}$) | 1.496 | 0.351 | 0.123 | 0.075 |
| FireNet+DeepPruner* ($\mathcal{L}_{gd}$ and $\mathcal{L}_{sm}$) | 1.336 | 0.355 | 0.123 | 0.068 |
| E2VID+DeepPruner* ($\mathcal{L}_{gd}$ and $\mathcal{L}_{sm}$) | 1.321 | 0.355 | 0.116 | 0.068 |
| FireNet+AANet (all losses) | 1.988 | 0.409 | 0.189 | 0.134 |
| E2VID+AANet (all losses) | 1.775 | 0.378 | 0.166 | 0.117 |
| FireNet+DeepPruner (all losses) | 1.626 | 0.377 | 0.147 | 0.097 |
| E2VID+DeepPruner (all losses) | 1.57 | 0.368 | 0.143 | 0.094 |
| FireNet+AANet* (all losses) | 1.201 | 0.306 | 0.110 | 0.065 |
| E2VID+AANet* (all losses) | 1.101 | 0.287 | 0.094 | 0.057 |
| FireNet+DeepPruner* (all losses) | 0.971 | 0.317 | 0.087 | 0.049 |
| E2VID+DeepPruner* (all losses) | 0.913 | 0.289 | 0.074 | 0.042 |

**Table II.** Quantitative comparison of different approaches on stereo matching using real-world dataset MVSEC [49]. ↑ means the higher the better while ↓ means the lower the better. - means the method completely fails.

| Model | EPE ↓ | Bad Pixels ↓ | | |
|---|---|---|---|---|
| | | $\delta > 1$ | $\delta > 3$ | $\delta > 5$ |
| Monodepth2 | 10.235 | 0.914 | 0.844 | 0.768 |
| E2VID+AANet (baseline) | 11.332 | 0.954 | 0.864 | 0.776 |
| E2VID+AANet (all losses) | 5.830 | 0.736 | 0.660 | 0.434 |
| E2VID+DeepPruner (all losses) | 4.979 | 0.673 | 0.581 | 0.384 |
| E2VID+AANet* (all losses) | 2.734 | 0.653 | 0.330 | 0.197 |
| E2VID+DeepPruner* (all losses) | 2.397 | 0.601 | 0.268 | 0.164 |

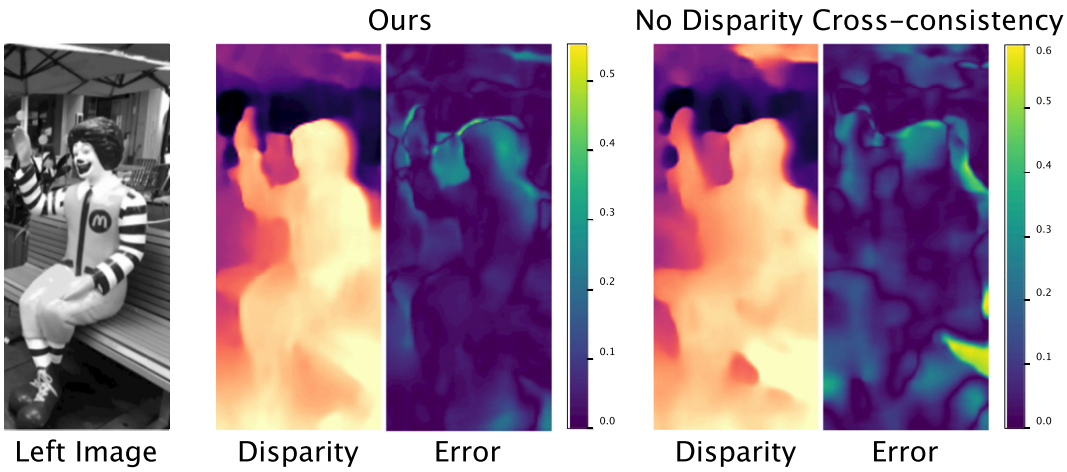


Figure 8. Comparison between our method with and without the cross-consistency loss. Our consistency term helps outline the precise edge in the disparity map.

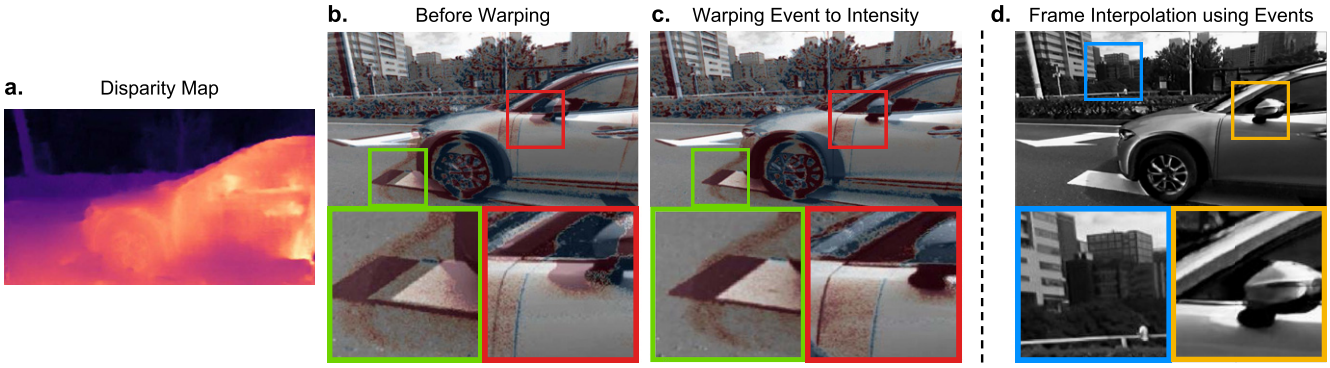


Figure 9. Connecting events and intensity. a, the calculated disparity map using the proposed method. b, before warping, events and intensity image are not aligned spatially. c, after warping using disparity map, the events and intensity image are well aligned. d, we employ the warped events and intensity image to perform temporal frame interpolation using [11].

image, it isn't easy to obtain a good result by employing the existing stereo vision models. The visualization results in Figure 5 also speak to similar conclusion. However, introducing the modified stereo network and self-supervised learning using only gradient structure loss can improve the stereo estimation results. Even if all the loss functions are used, the network architecture is still a significant obstacle to improving performance. Third, the proposed self-supervised learning method unleashes the full potential of the overall framework's effect, which proves the advantage of our self-supervised learning strategy, that is, learning from unlabeled data. We visualize the results of the proposed method in Fig. 5 and Figure 6. As one can observe, we can only get poor matching results in these scenarios based on the pre-trained stereo network. Our full model produces accurate object boundaries and better preserves the overall scene structures. We also provide the upper bound performance obtained by a fine-tuned DeepPruner network using both sides' intensity images as a reference. It can be seen that the information lost by the event is detrimental to the final matching result. But the purpose of our method is not to rely on events to obtain better matching results but to make it possible to calculate reasonable disparity under the proposed intensity-event setting.

### 4.3 *Ablation Study*

To study the effects of each component in the proposed method, we conduct several ablation studies. All the experiments are conducted using FireNet reconstruction and a modified AANet stereo model. We first examine the use of gradient structure loss function. We train our model using the four alternative loss functions described in Section 3.3 and their convergence curves are shown in the left figure of Figure 7. As can be seen, all pixel-wise loss functions fail to converge. They can only provide very limited information, and continuous optimization of these losses will bring adverse effects and make training fail. The structure loss directly calculated on images performs well initially, but it could no longer provide adequate supervision as the training progressed. The proposed gradient structural loss

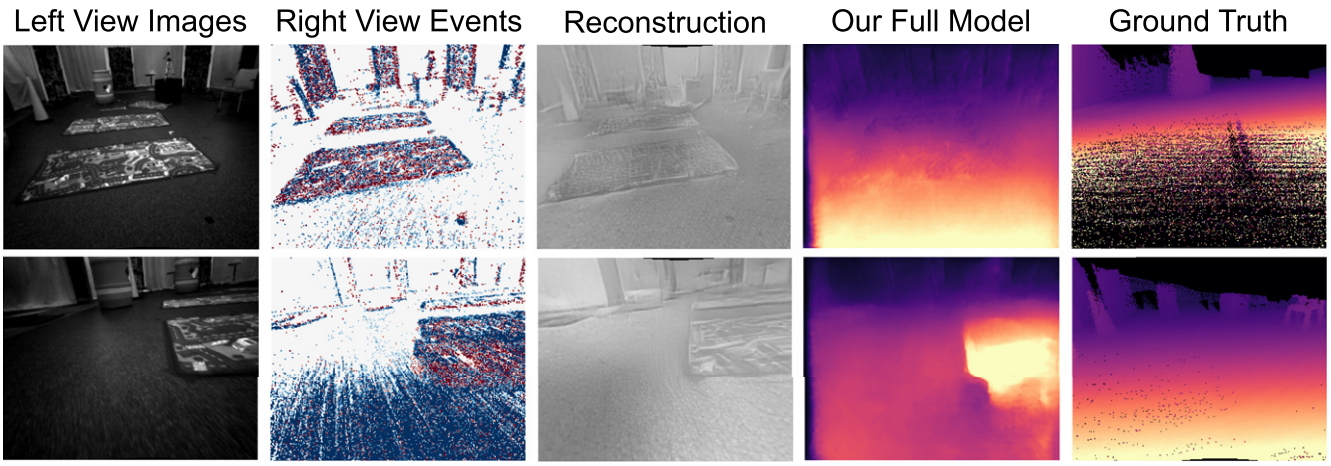


Figure 10. Failure cases visualization. The proposed method faces challenges when the reconstruction only contains very limited information.

can produce a relatively reliable convergence curve, although it is still tricky to improve performance with training steadily. We next involve the proposed cross-consistency loss and internal disparity loss in training. The results are shown in right figure of Fig. 7. It is surprising that although it cannot be compared with competitors initially, the proposed losses make the training process more stable. The proposed method can steadily and continuously improve performance without any ground truth data. We visually explain why the proposed cross-consistency loss is so effective. Figure 8 shows some comparisons between models trained with and without the cross-consistency loss. The results speak to the fact that the inclusion of this loss improves the quality of the result. This loss term helps outline the precise edge and generate sharp, accurate shapes in the disparity map.

### 4.4 *Connecting Events and Intensity*

We next demonstrate the use of a disparity map to establish a connection between the intensity camera and the event camera. Due to the displacement between these two cameras, even if we can obtain the left view intensity image and the right view events, the disparity between them makes it difficult to make full use of the advantages of the two sensors, as shown in Figure 9(b). In this case, many algorithms and applications that require alignment of events and images cannot be implemented, e.g., [11, 53]. We first obtain the disparity map using the proposed multi-modal stereo method. Each value in the disparity map indicates the number of pixels that need to be shifted horizontally. We warp events by changing the x coordinate in each event tuple $(x_m, y_m, t_m, p_m)$, where $x_m$, $y_m$, $t_m$ denote the spatial-temporal coordinates, and $p_m \in \{-1, +1\}$ denotes the polarity of the event. The warped event are visualized in Fig. 9(c). It can be seen that the warped right view events and the left view image are well aligned both spatially and temporally. Obtaining such a connection between two sensors allows many downstream tasks. Here we show the application potential by the event-based video interpolation task. Motivated by the physical model of event that the residuals between a blurry image and sharp frames are the integrals of events, Lin et al. [11] propose to estimate the residuals for the sharp frame restoration based on events. Our reconstruction result shows good fidelity performance, which further proves the application value of the proposed problem setting.

### 4.5 *Limitations*

Finally, we show some failure cases and analyze the potential limitations. We show two failure cases in Figure 10 and both of them are from MVSEC. The direct reconstruction using E2VID contains only very limited information, resulting in the failure of stereo matching. This shows a possible flaw of the proposed method, that is, it still faces significant challenges when reconstruction results are very vague. A possible solution is to design a stereo network to perform stereo matching between event streams and intensity images directly. In that case, the loss functions described in this paper can still provide good self-supervised learning results.

## 5. CONCLUSION

This paper presents a novel camera setting with an intensity camera and an event camera and establishes a connection between them with a multi-modal stereo matching task. Based on the proposed self-supervised method, we can obtain fine disparity maps under this novel setting and not collect any ground truth disparities. Experiments demonstrate the effectiveness and the application value of the proposed method.

## APPENDIX A. MORE STEREO MATCHING RESULTS

We first provide more stereo matching results using different methods. In Figure C4, we provide the comparison of synthetic events data. The synthesis method is described in Sec 4.1. In Figure C5, we provide the comparison results on the MVSEC [49] dataset. In these experiments, we use E2VID as the reconstruction network and AANet [30] as the stereo matching network. It can be seen that our method provides

the best results and performs well in these cases. In Fig. C5, we also compare our method with another commonly used method when encountering multi-modal problems, e.g., Zhi et al. [32] proposed to use a spectral translation network to facilitate cross-spectral stereo matching. We develop a similar adaptation network to translate the rough reconstructed right view image to add color information. The network structure is ResNet image translation structure, similar to SRResNet [54] but without an upsampling layer. We use a supervised training method to train this network, and the ground truth of the intensity image is provided in MVSEC. As can be seen, simply using the modality alignment method cannot bring better results. There are two main reasons for this. First, the color information has been lost in the events and reconstruction results and cannot be simply recovered by an adaptation network. Second, the pre-trained stereo model cannot generalize well in the MVSEC dataset. This also provides the necessities of learning using target data in a self-supervised manner.

## APPENDIX B. EVENT WARPING AND VIDEO FRAME INTERPOLATION

In this section, we show more results of warping events to intensity image. We first obtain the disparity between these two sensors using the proposed method. The warping provides us with events and intensity images that are aligned both spatially and temporally. We can use the obtained events and images to support downstream applications. In this supplementary material, we show more video frame interpolation [11] results in Figure C6.

The proposed method has additional value in this respect. Due to the hardware limitations of the event camera, high-resolution events and high-resolution intensity images are not available simultaneously. But with the proposed method, we can obtain high-resolution events and images simultaneously through two sensors. This makes a series of applications, such as video frame interpolation possible.

## APPENDIX C. EFFECT OF THE LOSS FUNCTIONS

We provide more results for the proposed general multi-modal stereo loss functions. We first show the convergence curves with different loss functions and metrics in Figure C1. It can be seen that the proposed loss functions enable the model to improve its performance through self-supervised learning continuously. The model without the proposed losses only provides a good guide initially, but when optimizing continuously, it does not match the purpose we want to achieve. Since the gradient structure loss will be numerically unstable when the difference between the two images is large, some methods will fail halfway. Although the internal disparity loss only provides simple, naive supervision, it has successfully improved the performance. We will further understand these two loss functions through visualization results. Figure C2 and Figure C3 show some error map visualization results. One can first observe from Fig. C3 that the proposed cross-consistency loss helps

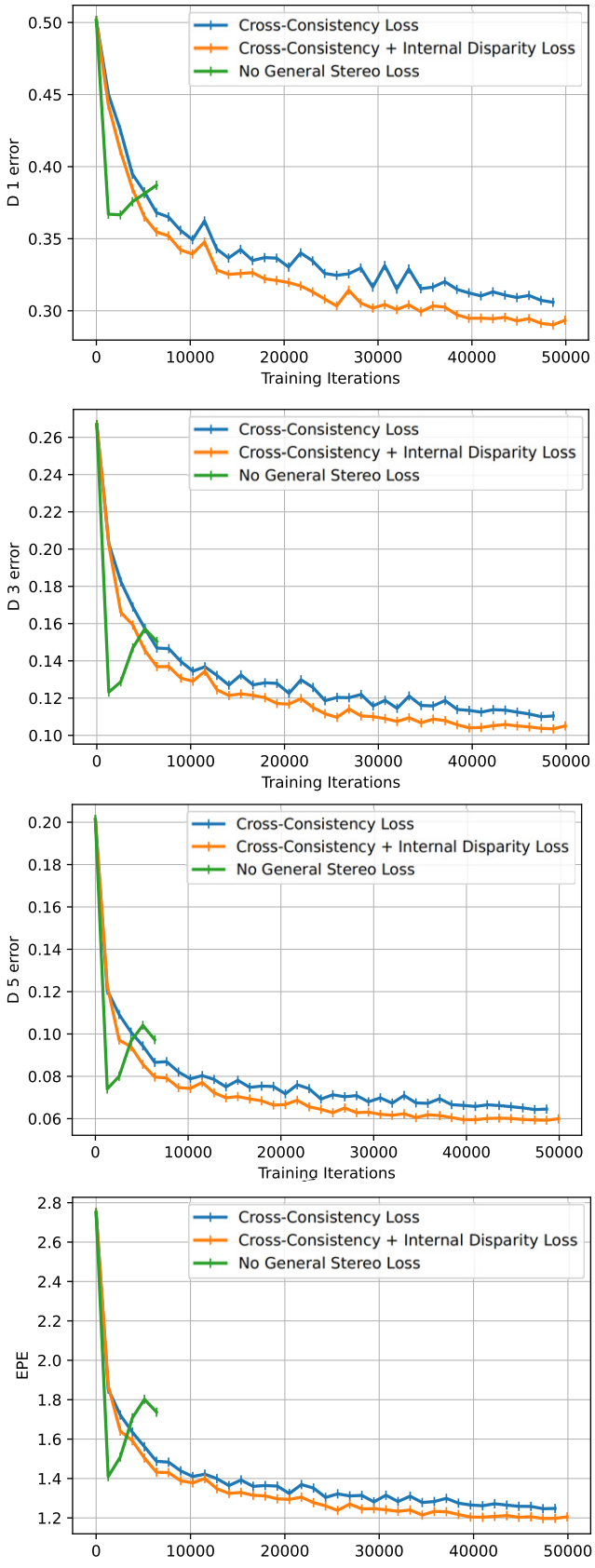


Figure C1. The convergence curves of using different losses. The proposed general stereo loss could produce robust training effects. The method of "no general stereo losses" failed in the middle of training.

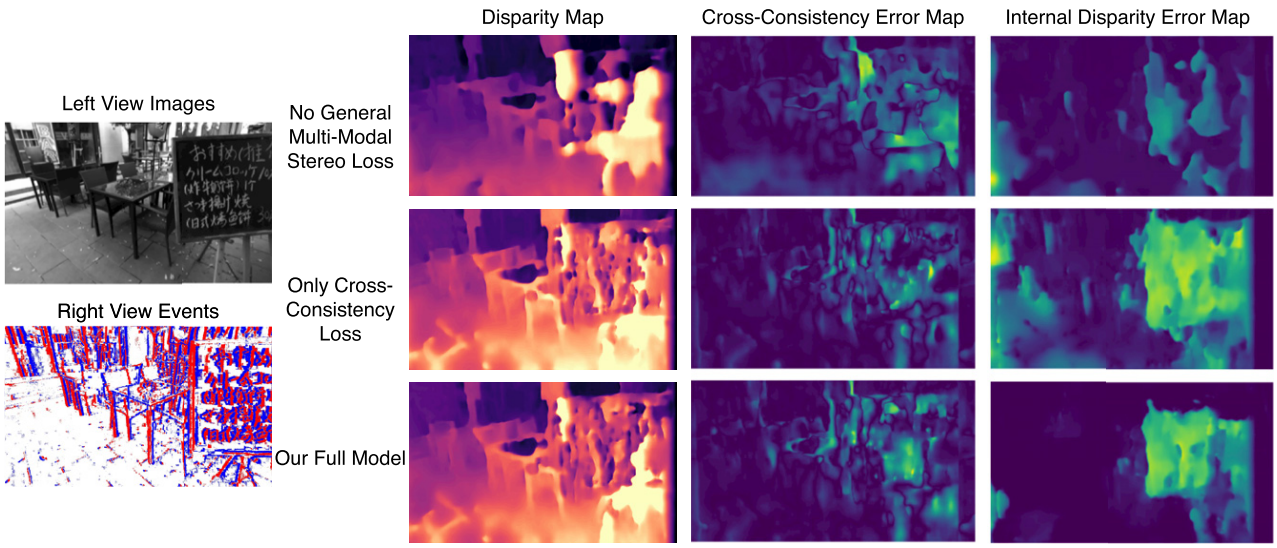


Figure C2. Visualization results of the proposed loss functions. This is a failure case. The disparity of the front object exceeds the upper limit of the network (in this case, the upper limit is 41 pixels). One can see that the proposed internal disparity loss points out where the error occurred.

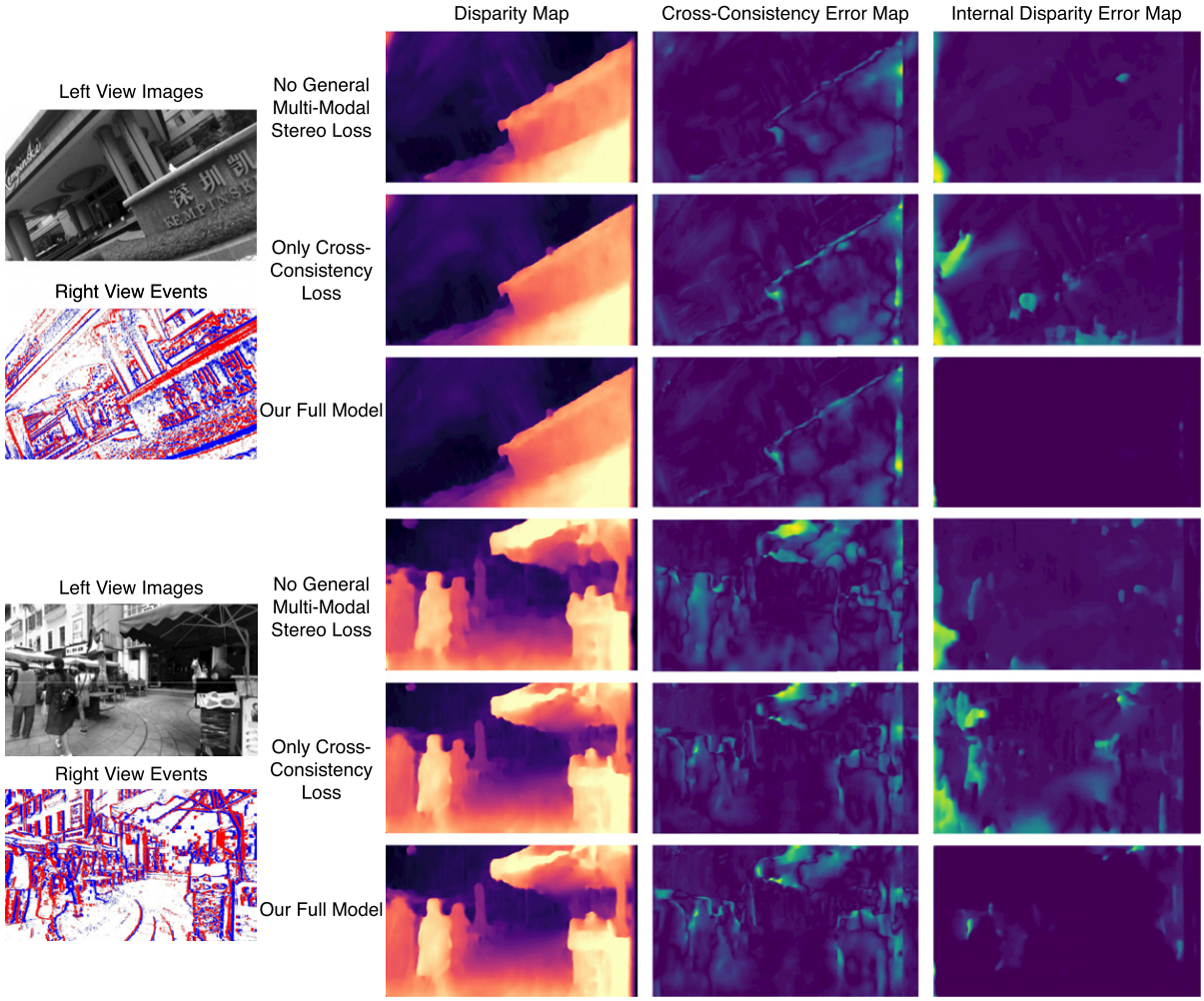


Figure C3. Visualization results of the proposed loss functions.

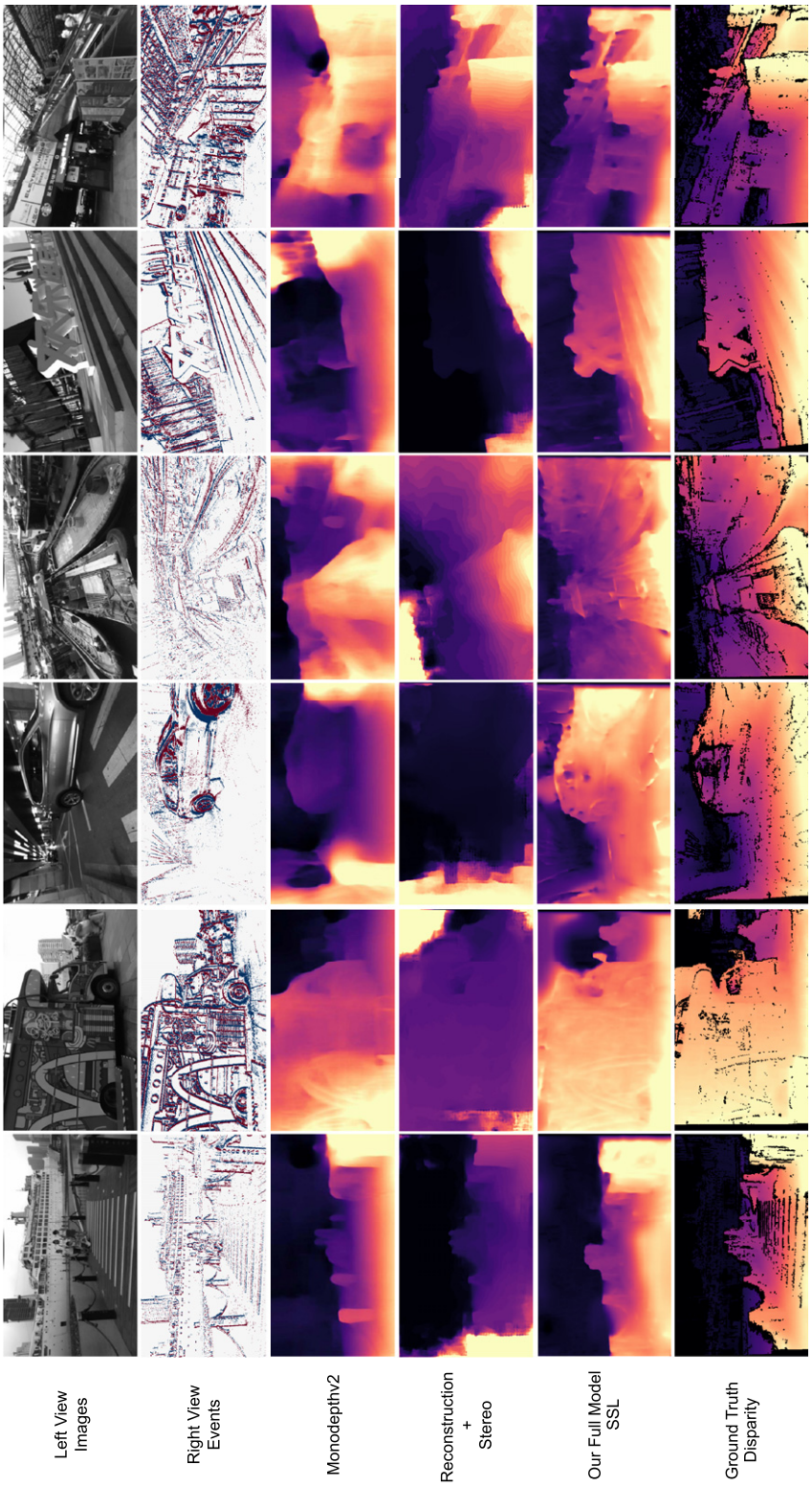


Figure C4. Stereo matching results of different methods on synthetic dataset.

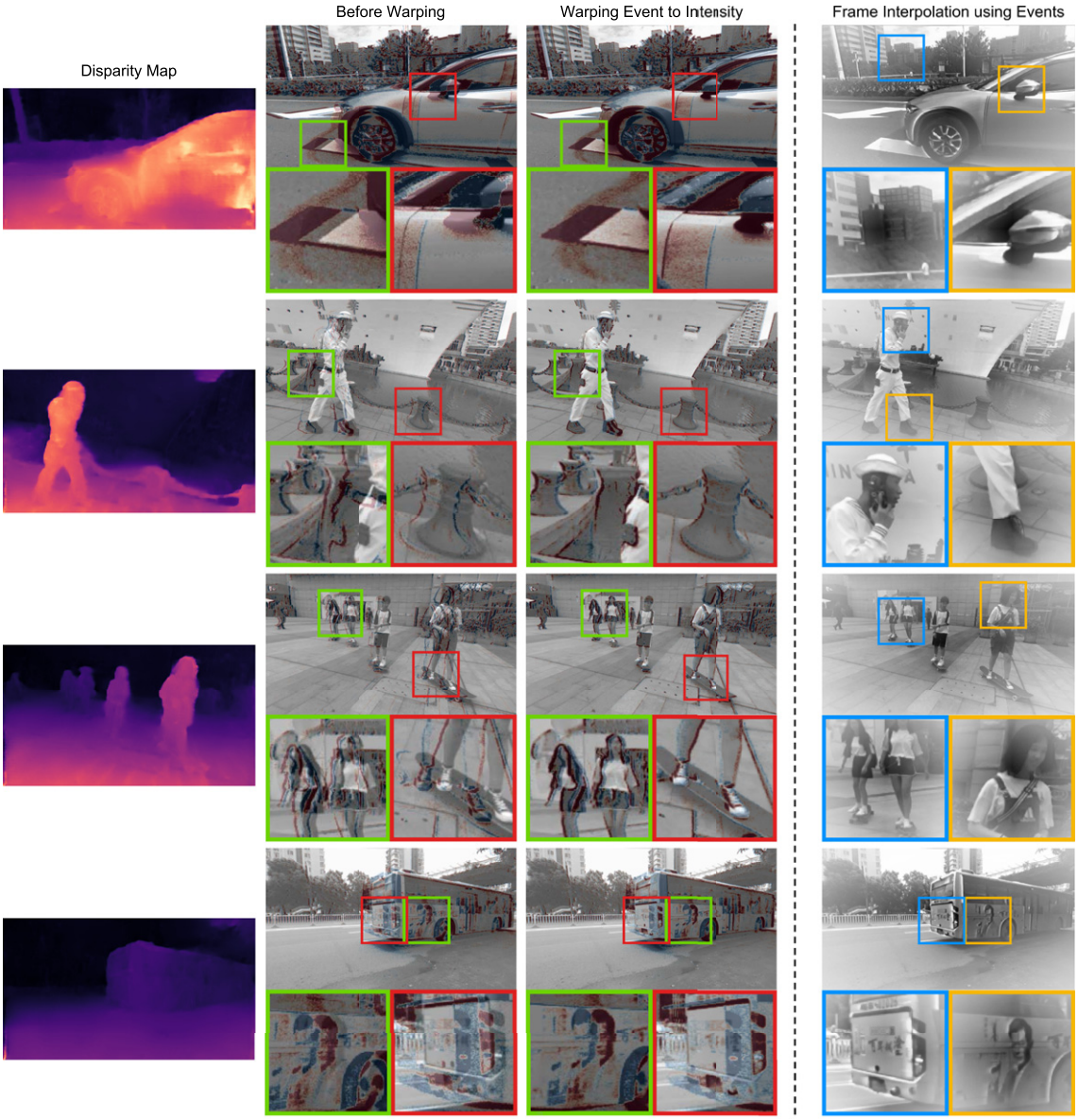


Figure C5. Stereo matching comparison of different methods on the MVSEC [49] real-world dataset.

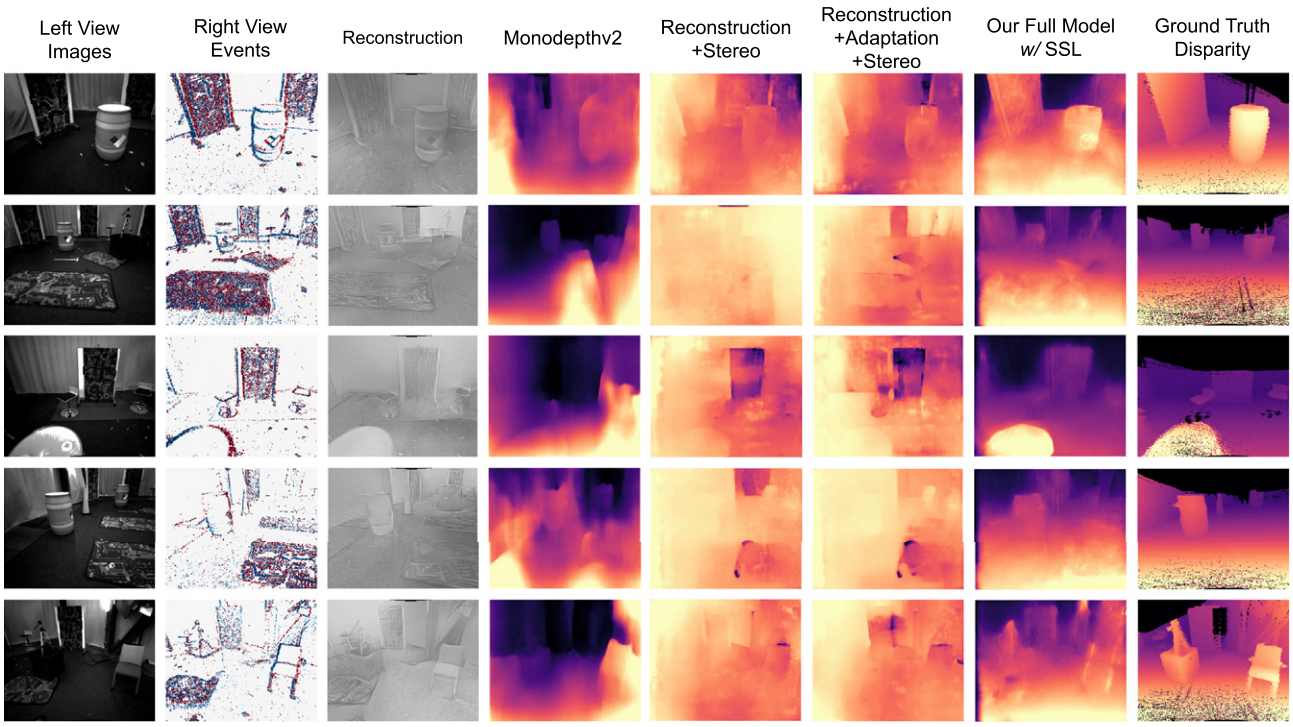


Figure C6. Connecting events and intensity. The event warping results and the temporal frame interpolation results using the warped events and intensity images using [11].

outline the edge and shape of the disparity. The proposed cross-consistency loss promotes the consistency of shapes between different views and provides additional information for training. We can also see from the cross-consistency error maps and internal disparity error maps that the introduction of these losses reduces the degree of these inconsistencies, especially for the internal disparity loss. Fig. C2 shows a failure case and also show how the internal disparity loss works. In this case, the disparity of the front object exceeds the upper limit of the network (we set the max disparity to be 41 pixels). The internal disparity loss reveals the failure area.

## APPENDIX D. ALTERNATIVE FRAMEWORK

We also present an alternative framework where the stereo network takes the right event voxel and the left intensity image as input directly. The framework is shown in Figure D1. In this framework, the network can also be the convolutional multi-modal stereo network. However, using convolution to process event voxel directly tends to bring poor results. We did not use this alternative since existing

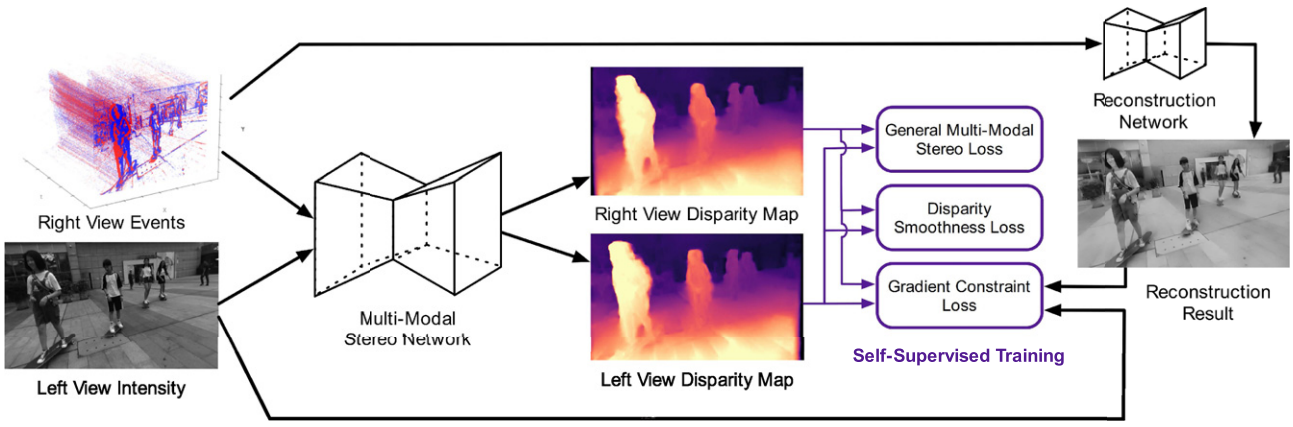


Figure D1. An alternative framework to the proposed self-supervised learning method. The stereo network takes the right event voxel and the left intensity image as input directly.

convolutional networks would be significantly better at processing images. But this alternative shows that our general multi-modal stereo consistency loss can be generalized to a wider range of application scenarios.

## REFERENCES


[1] H. Rebecq, T. Horstschäfer, G. Gallego, and D. Scaramuzza, "Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real time," IEEE Robot. Autom. Lett. **2**, 593–600 (2016).

[2] Z. Jiang, Y. Zhang, D. Zou, J. Ren, J. Lv, and Y. Liu, "Learning event-based motion deblurring," *CVPR* (IEEE, Piscataway, NJ, 2020), pp. 3320–3329.

[3] A. Z. Zhu and L. Yuan, "Ev-flownet: Self-supervised optical flow estimation for event-based cameras," Robotics: Science and Systems (2018).

[4] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios," IEEE Roboti. Autom. Lett. **3**, 994–1001 (2018).

[5] S. Tulyakov, D. Gehrig, S. Georgoulis, J. Erbach, M. Gehrig, Y. Li, and D. Scaramuzza, "Time lens: Event-based video frame interpolation," *CVPR* (IEEE, Piscataway, NJ, 2021), pp. 16155–16164.

[6] L. Patrick, C. Posch, and T. Delbruck, "A 128 × 128 120 db 15 µs latency asynchronous temporal contrast vision sensor," IEEE J. Solid-State Circuits **43**, 566–576 (2008).

[7] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, "A 240×180 130 db 3 µs latency global shutter spatiotemporal vision sensor," IEEE J. Solid-State Circuits **49**, 2333–2341 (2014).

[8] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-video: Bringing modern computer vision to event cameras," *CVPR* (IEEE, Piscataway, NJ, 2019), pp. 3857–3866.

[9] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. Mahony, and D. Scaramuzza, "Fast image reconstruction with an event camera," *WACV* (IEEE, Piscataway, NJ, 2020), pp. 156–163.

[10] D. Gehrig, H. Rebecq, G. Gallego, and D. Scaramuzza, "Asynchronous, photometric feature tracking using events and frames," *ECCV* (Springer, Cham, 2018), pp. 750–765.

[11] S. Lin, J. Zhang, J. Pan, Z. Jiang, D. Zou, Y. Wang, J. Chen, and J. Ren, "Learning event-driven video deblurring and interpolation," *ECCV* (Springer, Cham, 2020), pp. 695–710.

[12] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion," *CVPR* (IEEE, Piscataway, NJ, 2019).

[13] H. Kim, S. Leutenegger, and A. J. Davison, "Real-time 3d reconstruction and 6-dof tracking with an event camera," *ECCV* (Springer, Cham, 2016), pp. 349–364.

[14] B. Kueng, E. Mueggler, G. Gallego, and D. Scaramuzza, "Low-latency visual odometry using event-based feature tracks," *IROS* (IEEE, Piscataway, NJ, 2016), pp. 16–23.

[15] S. Barua, Y. Miyatani, and A. Veeraraghavan, "Direct face detection and video reconstruction from event cameras," *WACV* (IEEE, Piscataway, NJ, 2016).

[16] P. Bardow, A. J. Davison, and S. Leutenegger, "Simultaneous optical flow and intensity estimation from an event camera," *CVPR* (IEEE, Piscataway, NJ, 2016).

[17] L. Wang, I. S. Mohammad Mostafavi, Y.-S. Ho, and K.-J. Yoon, "Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks," *CVPR* (IEEE, Piscataway, NJ, 2019).

[18] I. S. Mohammad Mostafavi, J. Choi, and K.-J. Yoon, "Learning to super resolve intensity images from events," *CVPR* (IEEE, Piscataway, NJ, 2020).

[19] L. Wang, T.-K. Kim, and K.-J. Yoon, "EventSR: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning," *CVPR* (IEEE, Piscataway, NJ, 2020).

[20] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, N. Barnes, L. Kleeman, and R. Mahony, "Reducing the sim-to-real gap for event cameras," *ECCV* (Springer, Cham, 2020).

[21] K.-J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (IEEE, Piscataway, NJ, 2006), Vol. 28, pp. 650–656.

[22] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *IEEE TPAMI* (IEEE, Piscataway, NJ, 2013).

[23] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," *CVPR* (IEEE, Piscataway, NJ, 2015).

[24] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," *CVPR* (IEEE, Piscataway, NJ, 2016), pp. 5695–5703.

[25] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," *CVPR* (IEEE, Piscataway, NJ, 2016), pp. 4040–4048.

[26] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," *ICCV* (IEEE, Piscataway, NJ, 2017).

[27] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," *CVPR* (IEEE, Piscataway, NJ, 2018).

[28] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," *CVPR* (IEEE, Piscataway, NJ, 2019).

[29] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun, "DeepPruner: Learning efficient stereo matching via differentiable patchmatch," *ICCV* (IEEE, Piscataway, NJ, 2019), pp. 4384–4393.

[30] H. Xu and J. Zhang, "AANET: Adaptive aggregation network for efficient stereo matching," *CVPR* (IEEE, Piscataway, NJ, 2020), pp. 1959–1968.

[31] W. W.-C. Chiu, U. Blanke, and M. Fritz, "Improving the kinect by cross-modal stereo," *BMVC* (Dundee, 2011), pp. 116–110.

[32] T. Zhi, B. R. Pires, M. Hebert, and S. G. Narasimhan, "Deep material-aware cross-spectral stereo matching," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2018), pp. 1916–1925.

[33] X. Shen, L. Xu, Q. Zhang, and J. Jia, "Multi-modal and multi-spectral registration for natural images," *ECCV* (Springer, Cham, 2014), pp. 309–324.

[34] H.-G. Jeon, J.-Y. Lee, S. Im, H. Ha, and I. S. Kweon, "Stereo matching with color and monochrome cameras in low-light conditions," *CVPR* (IEEE, Piscataway, NJ, 2016), pp. 4086–4094.

[35] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn, "DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence," *CVPR* (IEEE, Piscataway, NJ, 2015), pp. 2103–2112.

[36] S. Kim, D. Min, S. Lin, and K. Sohn, "Deep self-correlation descriptor for dense cross-modal correspondence," *European Conf. on Computer Vision* (Springer, Cham, 2016), pp. 679–695.

[37] I. S. Mohammad Mostafavi, K.- J. Yoon, and J. Choi, "Event-intensity stereo: Estimating depth by the best of both worlds," *ICCV* (IEEE, Piscataway, NJ, 2021), pp. 4258–4267.

[38] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," *ECCV* (Springer, Cham, 2016), pp. 740–756.

[39] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," *CVPR* (IEEE, Piscataway, NJ, 2017).

[40] C. Godard, O. M. Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," *ICCV* (IEEE, Piscataway, NJ, 2019), pp. 3828–3838.

[41] C. Zhou, H. Zhang, X. Shen, and J. Jia, "Unsupervised learning of stereo matching," *ICCV* (IEEE, Piscataway, NJ, 2017), pp. 1567–1575.

[42] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (IEEE, Piscataway, NJ, 2021), Vol. 43, pp. 1964–1980.

[43] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *CVPR* (IEEE, Piscataway, NJ, 2018), pp. 586–595.

[44] G. Jinjin, C. Haoming, C. Haoyu, Y. Xiaoxing, J. S. Ren, and D. Chao, "Pipal: a large-scale image quality assessment dataset for perceptual image restoration," *ECCV* (Springer, Cham, 2020), pp. 633–651.

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process. **13**, 600–612 (2004).

[46] S. Zhou, J. Zhang, W. Zuo, H. Xie, J. Pan, and J. S. Ren, "DAVANet: Stereo deblurring with view aggregation," *CVPR* (IEEE, Piscataway, NJ, 2019), pp. 10996–11005.

[47] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super SloMo: High quality estimation of multiple intermediate frames for video interpolation," *CVPR* (IEEE, Piscataway, NJ, 2018), pp. 9000–9008.

[48] T. Delbruck, Y. Hu, and Z. He, "V2E: From video frames to realistic DVS event camera streams," *CVPR* (IEEE, Piscataway, NJ, 2021).

[49] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3d perception," IEEE Robot. Autom. Lett. **3**, 2032–2039 (2018).

[50] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "Pytorch: An imperative style, high-performance deep learning library," *NeurIPS* (Curran Associates, Red Hook, NY, 2019), pp. 8026–8037.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR* (ICLR, New Orleans, LA, 2015).

[52] D. Gehrig, M. Rüegg, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction," IEEE Robot. Autom. Lett. **6**, 2822–2829 (2021).

[53] P. Duan, Z. W. Wang, B. Shi, O. Cossairt, T. Huang, and A. Katsaggelos, "Guided event filtering: Synergy between intensity images and neuromorphic events for high performance imaging," IEEE Trans. Pattern Anal. Mach. Intell. **44**, 8261–8275 (2021).

[54] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," *ECCV* (Springer, Cham, 2018), pp. 63–79.