Learning Adaptive Dense Event Stereo from the Image Domain

Hoonhee Cho, Jegyeong Cho, and Kuk-Jin Yoon
Visual Intelligence Lab., KAIST, Korea
{gnsgnsgml, j2k0618, kjyoon}@kaist.ac.kr

Abstract

Recently, event-based stereo matching has been studied due to its robustness in poor light conditions. However, existing event-based stereo networks suffer severe performance degradation when domains shift. Unsupervised domain adaptation (UDA) aims at resolving this problem without using the target domain ground-truth. However, traditional UDA still needs the input event data with ground-truth in the source domain, which is more challenging and costly to obtain than image data. To tackle this issue, we propose a novel unsupervised domain Adaptive Dense Event Stereo (ADES), which resolves gaps between the different domains and input modalities. The proposed ADES framework adapts event-based stereo networks from abundant image datasets with ground-truth on the source domain to event datasets without ground-truth on the target domain, which is a more practical setup. First, we propose a self-supervision module that trains the network on the target domain through image reconstruction, while an artifact prediction network trained on the source domain assists in removing intermittent artifacts in the reconstructed image. Secondly, we utilize the feature-level normalization scheme to align the extracted features along the epipolar line. Finally, we present the motion-invariant consistency module to impose the consistent output between the perturbed motion. Our experiments demonstrate that our approach achieves remarkable results in the adaptation ability of event-based stereo matching from the image domain.

1. Introduction

Stereo matching [22, 41] is one of the most widely used methods for obtaining 3D information by establishing correspondences between stereo images. With considerable interest, learning-based stereo methods have achieved state-of-the-art performance in many benchmark datasets. However, some challenges in stereo matching still exist due to the shortcoming of sensors (*e.g.*, low dynamic range, motion blur due to large exposure time). Event cameras [3] are novel sensors that asynchronously report per-pixel changes of intensity by imitating the human eye. Thanks to the high dynamic range and low latency, the event camera can be considered as a promising sensor for depth estimation, especially in driving scenarios. Recent works [2, 8, 9, 28, 30, 48, 59] have attempted to utilize event cameras for stereo matching even under poor light conditions.

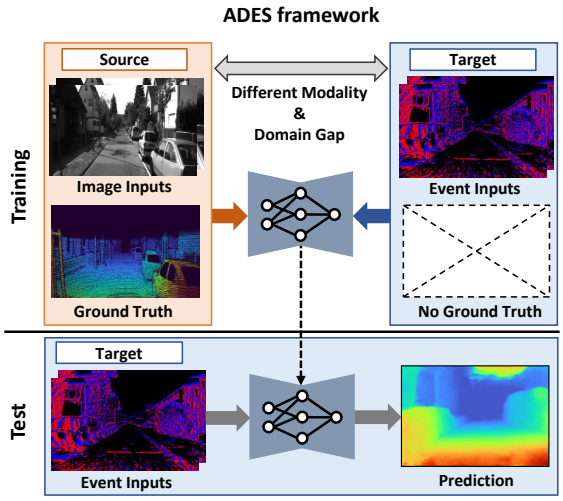

Figure 1. The proposed ADES framework for adaptive dense event stereo network. ADES aims to exploit the existing frame-based stereo dataset for learning the event stereo network.

Despite advances in event stereo, most prior works [9, 28, 48] still experience a significant degradation in performance when domains shift. Unsupervised domain adaptation (UDA) can resolve this problem without using the target domain ground-truth. When UDA is applied for event stereo domain adaptation, it still needs the input event data with ground-truth in the source domain. However, as mentioned in [32], accurate synchronization of events with high temporal resolution and other devices (*e.g.*, LiDAR) requires additional hardware and post-processing, so it is more challenging to obtain accurate ground truth than images. In this paper, we draw attention to large image datasets with ground-truth, which are easily accessible (*e.g.*, DrivingStereo [57], SceneFlow [24] and KITTI [25]). In this setup, abundant image data from diverse environments helps the event stereo network improve generalizability with high performance. To this end, as shown in

Fig. 1, we propose a novel Adaptive Dense Event Stereo (ADES), which adapts the stereo network from the source domain having image data with ground-truth to the target domain having event data without ground-truth. ADES resolves gaps between the different domains and input modalities.

The proposed ADES framework consists of three components: smudge-aware self-supervision module, feature normalization, and motion-invariant consistency module. The proposed smudge-aware self-supervision module leverages dense traits of images via image reconstruction on the event target domain. Image reconstruction using only the event is often interrupted by blurry artifacts, what we call a smudge, so the network cannot estimate the sharp and accurate disparity map. To predict the smudge effect in the target domain, we design the self-supervision pipeline on the source image domain to estimate and suppress the artifact area in the reconstructed image on the target domain.

In addition, we exploit the feature normalization before generating the cost volume. Normalization scheme [29, 43, 49, 58] was generally used in the domain adaptation between the image modalities. However, due to the characteristics of the event, it is not efficient to normalize over the entire pixel area. Since most of the events are triggered around an edge of objects, some regions (*e.g.*, sky) have very sparse events. Therefore, vanilla normalization can mislead the values of features to shift to the values of the regions without events. While reducing the difference in features between the two domains, we apply a normalization along the epipolar line to take into account the characteristics of events and stereo matching.

Finally, we focus on the different motion of event cameras from the source and target domains, leading to a severe domain gap. Therefore, we present the motion-invariant consistency module to predict consistent disparity even if the camera motion changes to some extent. This module help the network to adapt the target domain and also reduces the gap from camera motion. To the best of our knowledge, our work is the first attempt to move from unpaired image domain to event domain for stereo matching. Our main contributions are summarized as below:

- Our work is the first that transfers the disparity estimation task from the rich image dataset with ground-truth to the event stream, resolving gaps between the different domains and input modalities.
- We propose a novel adaptive event stereo network, ADES, containing the smudge-aware self-supervision module, feature normalization, and motion-invariant consistency module.
- Extensive experiments demonstrate that the ADES framework achieves significantly better performance than the prior works in the adaptation ability between the different domains and modalities for event stereo.

2. Related Works

2.1. Stereo depth estimation using Events

Recent event-based stereo matching [2,8,9,28,30,48,59] achieved the high accuracy than early works [5,7,20,35–37, 39, 60, 64, 65] by adopting a learning-based approach with various embedding scheme. They propose additional modules that consider the temporal continuity of events, but the overall framework is similar to frame-based stereo: embedding, matching, and regularization modules. Our proposed pipeline of adaptive stereo matching also follows the general stereo matching, then can be easily applied on other event-based stereo networks.

2.2. Domain Adaptation in Stereo Matching

To overcome the performance degradation from the domain gap, several works have explored unsupervised domain adaptation (UDA) in frame-based stereo matching. Some works utilize the knowledge distillation [14] and CycleGAN [23,62] to narrow the domain gap. After that, compact and efficient domain adaptation studies [33,43,47] for stereo have been conducted. On the other hand, the domain adaptation ability of event stereo networks has not been studied, and we tackle it for the first time.

2.3. Adaptation from Images to Events

Event-to-image reconstruction methods [10, 27, 31, 34, 38, 42, 44, 50, 52, 53] can be considered the proxy task to transfer a labeled image domain (source) to an unlabeled event domain (target). The results of the event-to-image reconstruction have been used as inputs of end-task (*e.g.*, object recognition, semantic segmentation) network pre-trained on image (source) domain. However, they introduce the extra latency in the inference time and still pose a domain gap between source and target domain.

One of the attempts to properly do domain adaptation was grafted networks [17] by utilizing the pre-trained image network. They replaced the encoder of pre-trained network with event encoder, then finetuned it on paired event and image datasets. This setup takes advantage of each modalities, but requires pixel-wise aligned events and images. To overcome the limitation of paired setting, EvDistill [51] leveraged the unpaired images to boost the performance of event-based networks. Through the bidirectional modality reconstruction and cross-modal knowledge distillation, they can transfer the knowledge from image network to event network. After that, research on transferring from an unpaired image to event has been of continuous interest. EV-Transfer [26] hallucinate the motion to generate the fake events from still images. ESS [46] achieve the feature-level alignment with the motion-invariant event embeddings.

Existing works only focus on the tasks from a single camera (*e.g.*, semantic segmentation, object recognition).

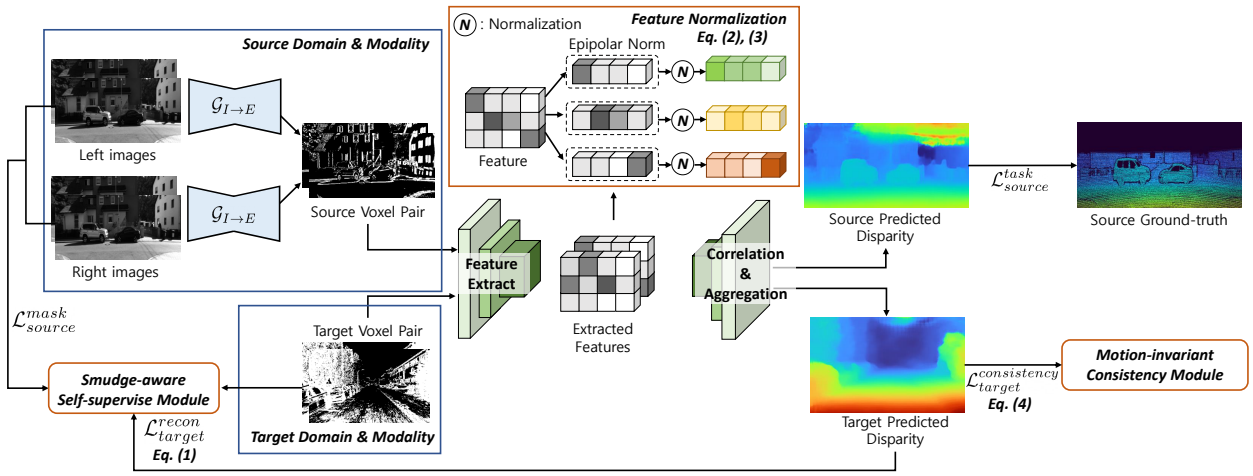

Figure 2. Overview of the proposed ADES framework. During training, image datasets (*e.g.*, KITTI) are used for the source domain, and event datasets without ground-truth are used for the target domain. We utilize feature normalization (Sec. 3.4) to narrow the gap between features from other domains. The proposed smudge-aware self-supervision module (Sec. 3.3) and motion-invariant consistency module (Sec. 3.5) are only used in training phase.

Due to the novel characteristics of event cameras, high-quality large stereo datasets have recently appeared, so related research has yet to be studied. In this paper, we propose an employable ADES framework for an event stereo. Unlike the prior works, we consider the characteristics of the event in the specific situation of stereo matching.

3. Method

3.1. Problem Setting

Given consecutive stereo image pairs (I_l^{t-1}, I_r^{t-1}), (I_l^t, I_r^t) with ground-truth disparity map $\tilde{d}_l^t$ on source domain, our goal is to train the model to predict the disparity map $D_l^{\hat{t}}$ at time $\hat{t}$ from stereo event streams $E_l^{\hat{t}}, E_r^{\hat{t}}$ on target event domain. To make it clear that the source and the target domains are unpaired, we denote time for the source domain as t and the target domain as $\hat{t}$. We represent an event stream as a voxel grid [63], the most commonly used representation, *e.g.*, by converting $E_l^{\hat{t}}$ to $V_l^{\hat{t}}$.

3.2. System Overview

As shown in Fig. 2, our proposed Adaptive Dense Event Stereo (ADES) consists of three novel components: smudge-aware self-supervision module, feature normalization, and motion-invariant consistency module.

For the source domain, we leverage the pre-trained video-to-event reconstruction network to extract the event representations from image data for the event-based stereo network. There are several video-to-event methods [11, 18], but among them, we adopt the network proposed in [61] that is lightweight and describes events well even with only two sequential images. Each left and right sequential image pairs are passed through the video-to-event network $\mathcal{G}_{I\to E}$ and transformed into voxel grids as follows: $V_l^t = \mathcal{G}_{I\to E}(I_l^{t-1}, I_l^t)$, $V_r^t = \mathcal{G}_{I\to E}(I_r^{t-1}, I_r^t)$. The generated voxel grid pairs (V_l^t, V_r^t) from the source domain and the voxel grid pairs $(V_l^{\hat{t}}, V_r^{\hat{t}})$ from the target domain are simultaneously fed into a weight shared event-based stereo network. In the process, we narrow the gap between the two domains by normalizing the extracted features. The prediction of source domain d_l^t is supervised by ground-truth disparity $\tilde{d}_l^t$, while the result of the target domain is adaptively optimized by the proposed smudge-aware self-supervision module and motion-invariant consistency module.

3.3. Smudge-aware Self-supervision Module (SSM)

Prior works [21, 45, 56] verified that self-supervised auxiliary tasks are helpful for domain adaptation, especially in stereo matching task [43]. We also utilize the photometric reconstruction as an auxiliary task of domain adaptation for stereo matching. First, as shown in the bottom of Fig. 3, we generate the image from voxel event representation in target domain. To this end, we leverage the widely used pre-trained event-to-image reconstruction network [38]. However, the reconstructed image from event has intermittent artifacts near the boundary of an object. Those blurring and distortion effects, what we call a smudge, disturbs predicting sharp disparity. To estimate the irregular smudge in the target domain, we further design a self-supervision pipeline for smudge prediction in the source domain.

In the source domain, as shown in the top of Fig. 3, we perturb the random region of image with distortion and blur kernel for imitating smudge effects. To depict realistic smudge effects, we do not generate the smudge on the randomly selected rectangle region but on the regions parsed by superpixel algorithm [1]. Since the edges of superpix-

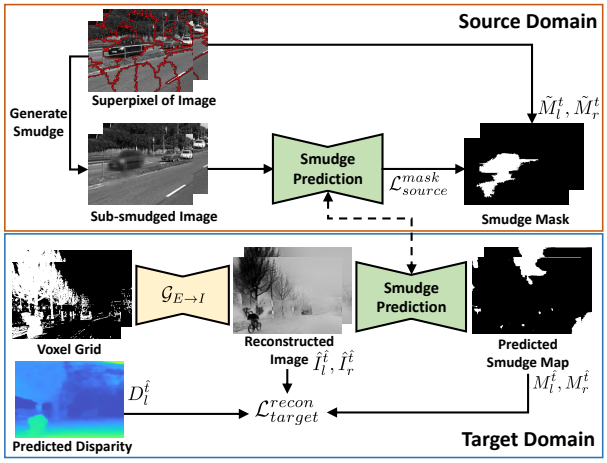

Figure 3. Illustration of the Smudge-aware Self-supervision Module (SSM). To estimate the smudge artifact of the reconstructed image on the target domain, we train the smudge prediction network to predict the smudge in the source domain.

els are generally placed on the boundary of an object, they can mimic the characteristics of the smudge effect caused at the boundary of the object due to the noisy sensors. Then, domain-shared smudge prediction network, which is a light-designed U-Net [40] with few convolution layers, is supervised from generated smudge mask. In other words, the output M_i^t of smudge prediction network is supervised from generated artifact mask $\tilde{M}_i^t$ by minimizing the binary cross entropy loss: $\mathcal{L}_{source}^{mask} = \sum_{i\in\{l,r\}} BCE(M_i^t, \tilde{M}_i^t)$.

In the target domain, as shown in the bottom of Fig. 3, we utilize a shared smudge prediction network from source domain for photometric-based self-supervision. The reconstructed left image $\hat{I}_l^{\hat{t}}$ and right image $\hat{I}_r^{\hat{t}}$ from voxels $V_l^{\hat{t}}$ and $V_r^{\hat{t}}$ are fed into the smudge prediction network. Then, smudge prediction network provides the predicted smudge map $M_l^{\hat{t}}$ and $M_r^{\hat{t}}$, which include per-pixel probability of smudge from 0 to 1. Next, from the predicted dense disparity map $D_l^{\hat{t}}$, we can reconstruct the warped left image $W_{r\to l}(\hat{I}_r^{\hat{t}})$ from right image $\hat{I}_r^{\hat{t}}$. Considering the smudge mask maps of both left and right sides, we can calculate the pixel-wise photometric reconstruction error as follows:

$$\mathcal{L}_{target}^{recon} = \alpha \frac{1 - \mathrm{SSIM}(\hat{I}_l^{\hat{t}} \odot M^{\hat{t}}, W_{r\to l}(\hat{I}_r^{\hat{t}}) \odot M^{\hat{t}})}{2} + (1-\alpha)\|\hat{I}_l^{\hat{t}} \odot M^{\hat{t}} - W_{r\to l}(\hat{I}_r^{\hat{t}}) \odot M^{\hat{t}}\|_1, \quad (1)$$

where $M^{\hat{t}} = 1 - (M_l^{\hat{t}} \odot W_{r\to l}(M_r^{\hat{t}}))$, $\odot$ means element-wise multiplication, SSIM denotes structural similarity proposed in [54] with a 3×3 kernel filter, and α is set to 0.85.

Our proposed smudge prediction network is shared in the source and target domains with the same weights. Compared to the stereo network, which is a high-level task, the smudge prediction, which pixel-wisely finds regions of noise and smudge, is a low-level task and less affected by

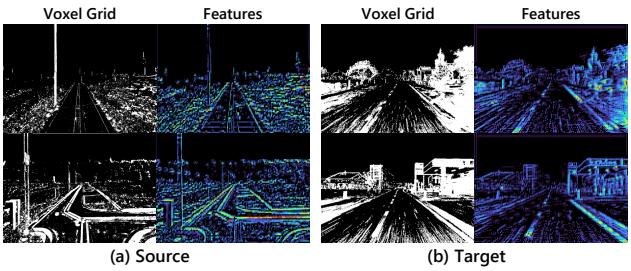

Figure 4. Visualization of voxel grids and features extracted by the event stereo network from source (*e.g.*, KITTI) and target (*e.g.*, DSEC) domain. In the upper region of features with few events, there are no activated pixels; in contrast, there are many activated pixels in the lower region with a lot of structure for events.

the domain gaps. Therefore, in the target domain, artifacts in the boundary are well predicted in the reconstructed image to estimate the sharp disparity through self-supervision.

3.4. Feature Normalization

We adopt the feature-level normalization to reduce the gap in the cost volume between two distinct domains. To align the distributions between different domains, the feature normalization [29,49,58] is widely used for image data. Similarly, event features can be also normalized to reduce the gap between the two domains; however, as can be seen in our experimental results (see Table 5), due to the sparsity of the event data, an existing normalization technique [43] for images rather hinders learning of event-based network.

For event feature normalization, we focus on the characteristics of event-based stereo matching. Events are naturally sparse and triggered predominantly to the edges of the objects. For accurate prediction with sparse information, event stereo networks tend to focus intensely on areas with event information, which leads to features of areas without events not being activated. Therefore, normalizing over all pixels provides results with shifted bias on the region without an event for each scene. Instead, we focus on the correlation between the amount of events and spatial positions. For example, events rarely occur in the upper regions (*e.g.*, sky), while more events occur in the lower regions where illuminance easily changes and rich textures exist (*e.g.*, building). This phenomenon can be qualitatively verified by the extracted features of the network, as shown in Fig. 4. The lower region of the feature map is strongly activated in both domains, while the upper region, where no event occurs, is hardly activated. Therefore, applying the normalization over all pixels used in the dense image [43] to event stereo rather hinders reducing the domain gaps. In addition, since the stereo disparity is defined as the position difference in the horizontal direction when the cameras are rectified, normalization over all spatial positions leads to ambiguity in accurate matching. Therefore, we propose a strategy to normalize the feature along the epipolar (hori-

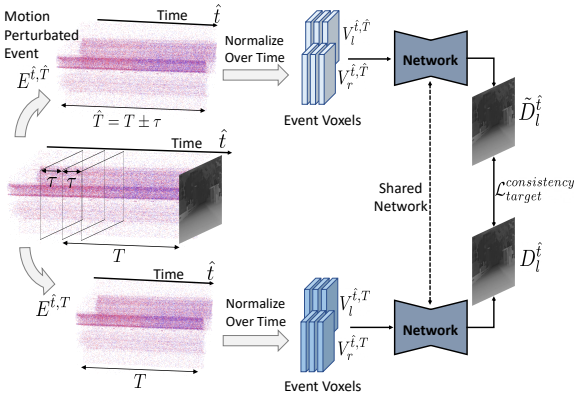

Figure 5. Illustration of the Motion-invariant Consistency Module (MCM). Normalize over time denotes to converting temporal values into a range between 0 and 1.

zontal) line, which does not violate the regularity according to regions. This normalization is also spatial but less dependent on the global distribution of event data.

We modify the parameter-free normalization scheme proposed in prior works [43]. Given extracted feature F from network with size of $C \times H \times W$ (C: channel, H: height, W: width), let $F(k,i,j)$ denotes kij-th element, where k is the index of channel and i, j are spatial dimensions. Then, we apply the normalization along channel-dimension to both left and right features as follows:

$$F(k,i,j) = \frac{F(k,i,j)}{\sqrt{\sum_{c=0}^{C-1} \|F(c,i,j)\|^2 + \varepsilon}}. \quad (2)$$

After that, we apply the normalization along the epipolar line, which is defined as:

$$F(k,i,j) = \frac{F(k,i,j)}{\sqrt{\sum_{w=0}^{W-1} \|F(k,i,w)\|^2 + \varepsilon}}. \quad (3)$$

3.5. Motion-invariant Consistency Module (MCM)

The motivation of the motion-invariant consistency module on the target domain is two-folds. The first is to resolve the domain gap caused by different camera motion. Second, we desire the event stereo network to be robust on perturbation or noise, adapting to the target domain. The perfectly adapted network should be able invariant to the motion of event camera and estimate the prediction stably, even if there is some perturbation in the input.

In this subsection, we denote the voxel grid corresponding to the events $E^{\hat{t},T}$ accumulated during time T until $\hat{t}$ as $V^{\hat{t},T}$. Specifically, events are converted to a voxel grid pair $V_l^{\hat{t},T}, V_r^{\hat{t},T}$, and fed to the network to estimate a disparity map $D_l^{\hat{t}}$. Since the motion in the datasets is already acquired and immutable, we use a trick to augment event streams of fast or slow motions with temporal perturbation τ. If events are stacked for a longer time $T + \tau$ (*i.e.*, more events) after normalizing temporal value in events (converting temporal values into a range between 0 and 1) and converted for the same bin of voxel grid, it is able to imitate the voxel grid of an event from an actual fast motion. Conversely, if stacked for a shorter time $T - \tau$ (*i.e.*, fewer events), it is the same as the voxel grid made with events of slow motion. From this, we can generate a motion perturbed event voxel $V^{\hat{t},\hat{T}}$, where $\hat{T} = T \pm \tau$ without additional data acquisition. As shown in Fig. 5, perturbed voxel pairs $V_l^{\hat{t},\hat{T}}, V_r^{\hat{t},\hat{T}}$ are fed into the event stereo network to predict the perturbed disparity map $\tilde{D}_l^{\hat{t}}$. Then, for pixel-wise consistency between the prediction of motion-perturbed input and the original input, our consistency loss is defined via L_1 distance:

$$\mathcal{L}_{target}^{consistency} = \|D_l^{\hat{t}} - \tilde{D}_l^{\hat{t}}\|_1 \quad (4)$$

More details about the implementation of motion-invariant consistency module can be seen in *supple*.

3.6. Loss functions

On the source domain, we use the two losses for training the network. We adopt the smooth L_1 loss for disparity estimation: $\mathcal{L}_{source}^{task} = \text{smooth}_{L_1}(\tilde{d}_l^t - d_l^t)$, and the binary cross entropy loss for artifact prediction: $\mathcal{L}_{source}^{mask} = \sum_{i \in \{l,r\}} BCE(M_i^t, \tilde{M}_i^t)$.

On the target domain, we use the reconstruction loss $\mathcal{L}_{target}^{recon}$ (Eq. 1) and consistency loss $\mathcal{L}_{target}^{consistency}$ (Eq. 4).

Finally, the total loss for an end-to-end optimization process is defined as:

$$\mathcal{L}^{total} = \mathcal{L}_{source}^{task} + \lambda_1 \mathcal{L}_{source}^{mask} + \lambda_2 \mathcal{L}_{target}^{recon} + \lambda_3 \mathcal{L}_{target}^{consistency}, \quad (5)$$

where λ_1, λ_2 and λ_3 are weight for each loss terms.

4. Experiments

4.1. Datasets

For source domain datasets, we utilize the KITTI dataset [25] and the SceneFlow dataset [24]. The KITTI dataset is a real-world dataset with two subsets (*i.e.* KITTI 2012 [13] and KITTI 2015 [25]), containing 394 stereo images with sparse ground-truth for training. On the other hand, the SceneFlow dataset [24] is a large synthetic dataset containing diverse scenes. The SceneFlow dataset provides 35k stereo images with dense ground-truth disparity maps.

For target domain, we use the recently published benchmark dataset of DSEC [12], a large-scale high-quality driving dataset with challenging scenes. DSEC provides high-resolution (640×480) stereo event streams captured in outdoor driving scenes. It contains 53 driving scenarios taken in various lighting conditions, and provides 17k stereo pairs

Table 1. Cross-domain comparisons with other traditional / domain generalization / domain adaptation stereo methods from various source domains. The 2-pixel error (%), 3-pixel error (%), end-point-error, and root mean square error are adopted for evaluation. Zu and In denote the Zurich City and Interlaken sequences, respectively.

Method	KITTI-to-DSEC								SceneFlow-to-DSEC							
	2PE		3PE		EPE		RMSE		2PE		3PE		EPE		RMSE	
	Zu	In	Zu	In	Zu	In	Zu	In	Zu	In	Zu	In	Zu	In	Zu	In
E2VID [38] on target domain																
SGM [16]	53.7	55.3	47.7	49.7	9.3	10.1	16.1	17.0	53.7	55.3	47.7	49.7	9.3	10.1	16.1	17.0
GwcNet [15]	38.2	37.1	29.1	22.0	3.5	2.9	8.2	5.8	43.3	45.0	28.8	31.2	3.0	4.2	6.5	12.2
PSMNet [6]	36.5	35.5	25.9	24.1	3.4	2.9	6.0	5.1	36.4	39.7	24.3	29.6	2.6	5.4	4.9	14.7
AANet [55]	46.3	39.3	35.3	26.3	7.4	3.8	17.2	8.6	42.8	39.2	29.2	27.3	3.6	3.1	7.2	5.1
DSMNet [58]	41.1	41.6	30.4	32.1	3.2	4.2	5.6	8.1	46.6	45.3	31.2	30.6	4.1	3.7	5.1	16.2
StereoGAN [23]	70.9	68.8	66.4	63.1	11.4	9.8	13.3	12.2	73.2	72.2	68.0	73.1	15.4	13.4	15.4	14.3
EventGAN [61] on source domain																
GwcNet [15]	25.5	27.3	19.8	16.8	4.6	3.1	12.2	9.1	55.1	56.3	37.2	37.1	7.4	7.3	16.4	16.1
PSMNet [6]	18.9	20.5	11.9	13.6	4.2	2.7	14.8	10.4	48.8	50.6	35.0	37.8	6.8	5.8	17.6	13.1
AANet [55]	50.4	47.3	42.5	38.6	9.8	5.4	21.7	11.3	58.8	55.4	45.3	42.4	5.7	5.8	11.0	10.6
DSMNet [58]	18.8	23.6	10.5	12.6	2.1	2.4	4.5	6.7	51.2	52.3	44.6	43.2	6.2	5.9	18.3	14.4
StereoGAN [23]	61.4	60.3	52.8	50.7	15.3	13.7	18.9	18.9	75.2	71.7	65.5	66.3	13.3	10.6	20.3	21.0
ADES (Ours)																
AANet	13.8	19.9	7.1	9.8	1.4	1.7	2.7	3.4	22.3	24.9	11.1	13.9	1.6	1.9	2.7	3.4
PSMNet	**10.9**	**10.2**	**5.6**	**5.5**	**1.2**	**1.3**	**2.5**	**3.2**	**17.5**	**12.8**	**9.4**	**6.7**	**1.4**	**1.3**	**2.7**	**2.8**

for training. Following the previous image-to-event transfer tasks for segmentation [46], we adopt sequences acquired during the day with monotonous illuminance, called *Zurich City*. Furthermore, we also conduct experiments on *Interlaken* sequences containing the high dynamic range and challenging illuminance scenes. More details about a split of datasets are provided in the *supple*.

4.2. Implementation details

The weights λ_1, λ_2 and λ_3 are set as 0.3, 1, and 0.2, respectively. We train our end-to-end framework using the Adam optimizer [19]. We set the learning rate to 1×10^{-3} with a batch size of 8 using 384×336 random crops. To generate the artifacts in Sec. 3.3, we utilize the Blur and OpticalDistortion transforms from library [4]. More details are provided in *supple*.

4.3. Comparisons with Other Methods

We compare our proposed ADES framework with the other traditional algorithm (SGM [16]), domain generalization (DSMNet [58]), and domain adaptation (StereoGAN [23]). In addition, we evaluate PSMNet [6], GwcNet [15], and AANet [55], which have achieved comparable performance in stereo matching. We train those networks in the source domain and test the network in the target domain. To align the different modalities between the source and target domains, we utilize E2VID [38] or EventGAN [61]. In the case of E2VID, an image-based stereo network is trained using a grayscale image in the source domain, and when testing, events in the target domain are reconstructed into a gray image using E2VID. Conversely, in the case of EventGAN, the event-based stereo network is trained using the events, which are converted by EventGAN from consecutive images in the source domain. We evaluate the proposed framework by applying it to AANet and PSMNet, which can represent networks using 3D and 4D cost volumes. Especially, AANet is efficient, so it is a widely adopted structure in recent event stereo research [28, 30]. When training, the proposed modules are used, but when evaluating, all the proposed modules without normalization are removed, and only the event stereo network is used to predict the disparity. Table 1 shows the results of training on the various source dataset, such as KITTI and SceneFlow datasets.

KITTI-to-DSEC. Even if the networks are trained on the source domain, the results of networks using the events from EventGAN can not achieve high performance. Some networks (PSMNet, GwcNet, and AANet) do not have the ability of domain generalization and adaptation. In addition, the distribution gap of disparity and motion between the domains lead to further degradation of performance. Similarly, most of the results using reconstructed images with E2VID do not show comparable performance. The reconstructed image on the target domain has a large gap with the real image on the source domain, which cannot be dealt with generalization (DSMNet), adaptation (StereoGAN) and conventional methods (SGM). Compared with existing works, our network achieves significantly high performance, considering the gap between domains such as motions and disparity. It is noticeable that in case of AANet+ADES, the performance improvement is significant, *e.g.*, 2PE decreases from 50.4 to 13.8, and RMSE decreases from 21.7 to 2.7 in Zurich City sequence.

SceneFlow-to-DSEC. The experiment results trained on the synthetic source domain show a more significant domain gap. Due to the effect of the syn-to-real gap, prior works show significantly lower performance than those

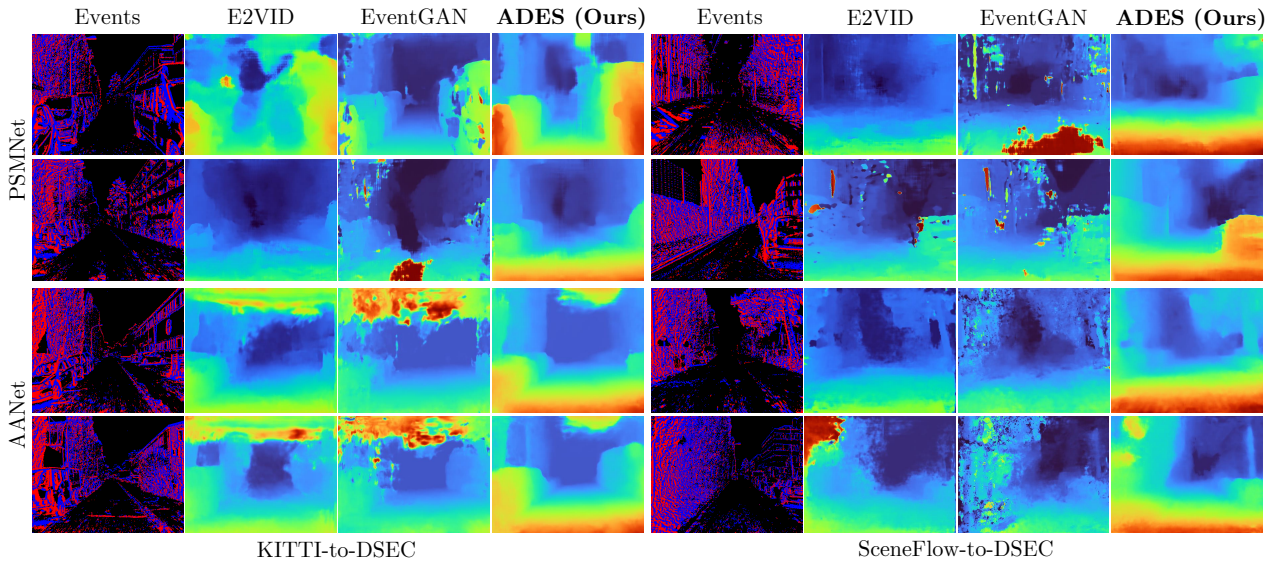

Figure 6. Qualitative results for the proposed method with other methods. Compared to EventGAN [61] and E2VID [38], our method can predict accurate and sharp disparity maps.

Table 2. Results on DSEC benchmark testset

Method	Supervision	RMSE ↓	MAE ↓	1PE ↓	2PE ↓
[48]	✓	1.386	0.576	10.92	2.91
[28]	✓	**1.222**	0.529	9.96	2.65
[30]	✓	1.231	**0.519**	9.58	2.62
[59]	✓	1.264	0.527	**9.52**	**2.36**
Ours (PSM)		1.698	0.771	18.37	5.36
Ours (AA)		1.982	0.936	24.01	7.87

trained on KITTI datasets, *e.g.*, 2PE of PSMNet using EventGAN significantly increases from 20.5 to 50.6 in the Interlaken sequence. Although our proposed method cannot avoid the performance degradation, the decreasing amount is relatively acceptable, *e.g.*, 2PE of PSMNet+ADES increases from 10.2 to 12.8 in the Interlaken sequence. Furthermore, the performance increases in some metrics thanks to the various scenes in the synthetic dataset, *e.g.*, RMSE of PSMNet+ADES decreases from 3.2 to 2.8 in the Interlaken sequence. These results show that our method can be used universally and not limited to a specific source dataset, whether synthetic or real-world.

In Fig. 6, we provide qualitative comparisons of our method with other methods. E2VID or EventGAN can reduce the gap between modality, but it cannot bridge between domains. On the other hand, our ADES framework predicts sharp and accurate disparity on both AANet and PSMNet. These results demonstrate that our training pipeline resolves the gaps from both domain and modality. More qualitative comparisons are provided in *supple*.

4.4. Results on DSEC test dataset

To compare with supervised methods, we utilize the KITTI dataset, which contains ground-truth, as the source domain and the full DSEC train dataset, which lacks ground-truth, as the target domain. Table 2 demonstrates that despite not relying on ground-truth data for the target domain, our approach achieves comparable performance to other supervised methods.

Table 3. Ablation studies for sub-modules: Smudge-aware Self-supervision Module (SSM), Feature Normalization (FN), and Motion-invariant Consistency Module (MCM). $D1$-error (%) is adopted.

Model	SSM	FN	MCM	Zurich	Interlaken
AANet				35.2	26.3
			✔	33.1	24.5
		✔		29.6	22.7
	✔			13.3	15.1
	✔		✔	11.2	14.6
	✔	✔		8.1	10.2
	✔	✔	✔	**7.0**	**9.7**
PSMNet				15.6	17.2
			✔	13.8	16.1
		✔		13.1	15.5
	✔			7.3	8.9
	✔		✔	6.6	7.1
	✔	✔		6.0	6.2
	✔	✔	✔	**5.6**	**5.7**

4.5. Ablation Studies

Effectiveness of each components. In Table 3, we conduct the ablation studies to validate the effectiveness of each component. We set the KITTI and DSEC dataset as source and target domain, respectively. Based on AANet and PSMNet from which all proposed modules were removed, the performance was measured while adding the proposed module one by one. The SSM achieves the greatest performance gain as a unitary component, *e.g.*, reducing error rates by 21.9% on Zurich, and 11.2% on Interlaken from AANet. In addition, using FN stand-alone reduces the

Table 4. Comparison of results with and without smudge prediction. $D1$-error (%) is adopted.

Methods	Zurich	Interlaken
w/o Smudge Prediction	16.1	12.3
w/ Smudge Prediction	**7.0**	**9.7**

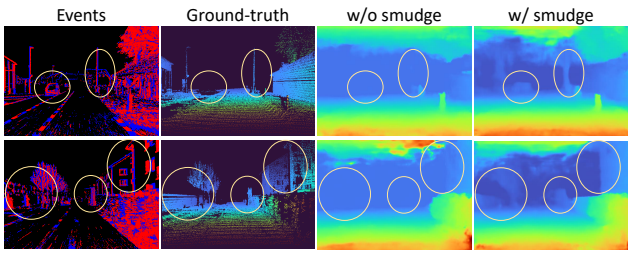

Figure 7. Qualitative results of disparity predictions with and without smudge prediction.

Table 5. Comparison of epipolar norm with existing cost norm [43]. $D1$-error (%) is adopted.

Methods	Zurich	Interlaken
w/o Norm	<u>11.2</u>	14.6
Cost Norm [43]	12.3	<u>11.1</u>
Our Epipolar Norm	**7.0**	**9.7**

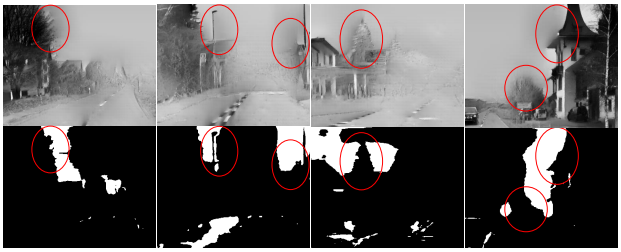

Figure 8. Generated smudge mask on the target domain. **Top:** reconstructed images, **Bottom:** corresponding smudge mask.

error by 1.7%∼5.6% depending on the target domain and model, compared to the baseline. Similarly, MCM also reduces the error by 1.1%∼2.1%. These FN and MCM work even if implemented together with SSM, *e.g.*, compared to the case of using only SSM from AANet in Zurich, when FN and MCM are used together, errors are more reduced by 5.2% and 2.1%, respectively. Finally, the performance gain is the most significant when all of the proposed modules are used, and these results validate that all modules are effective for domain adaptation.

Effectiveness of smudge predictions. We investigate the effect of the proposed smudge prediction on SSM. Using the KITTI as the source domain, the quantitative and qualitative results with and without smudge prediction from AANet are shown in Table 4 and Fig. 7. Compared with the absence of smudge prediction, our method utilizes the photometric loss of image reconstruction effectively. Therefore, through SSM using smudge prediction, the performance of the disparity estimation is improved *e.g.*, the $D1$-error decreases from 16.1 to 7.0 in Zurich City sequence. Specifically, as can be seen in Fig. 7, smudge prediction helps the network to robustly infer the boundary of an object, making a sharp disparity map. We visualize the smudge mask in Fig. 8 to show how smudge masks can aid effective learning in SSM. As shown in 1st row of Fig. 8, when an image is reconstructed from an event via E2VID, smudge-like blurry artifacts appear around the boundary of the object due to the noisy nature of the event streams. Our smudge prediction estimates these artifacts regions, helping to sharpen disparity when performing self-supervision and boosting the performance of disparity estimation.

Comparison of normalization. Table 5 reports the comparison of our feature normalization with the existing learning-free normalization method [43]. We set the KITTI dataset as the source domain and the model to AANet. Although our epipolar norm is a modified module of the existing work, the results verify that it is more effective for the domain adaptation ability of event stereo matching. For example, in the case of [43], which performs the normalization along all pixels in spatial manners, the performance is rather reduced due to the specificity of event stereo. For example, especially in Zurich City, the $D1$-error increases from 11.2 to 12.3 after introducing the cost norm [43]. On the other hand, our epipolar norm shows a performance improvement in both sequences. From these results, we can confirm that the event stereo network requires a different approach than the existing normalization method, and our normalization is suitable for the domain adaptive event stereo matching.

5. Conclusion

In this paper, we propose a novel framework, ADES, for adaptive dense event stereo from the image domain. Our work is the first that transfers the disparity estimation task from the rich image dataset with ground-truth to the event stream to tackle the absence of ground-truth disparities on the target event domain. To this end, we propose the smudge-aware self-supervision module, feature normalization, and motion-invariant consistency module by focusing on the specificity of event stereo. Extensive experiments demonstrate our framework achieves remarkable performance on both 3D and 4D cost volume networks, whether the source domain is synthetic or real-world image dataset. Furthermore, these results demonstrate that our modules can be used universally in existing event stereo networks. We open up the possibility of using event stereo in domains without ground-truth and expect more future work.

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

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 3

[2] Soikat Hasan Ahmed, Hae Woong Jang, SM Nadim Uddin, and Yong Ju Jung. Deep event stereo leveraged by event-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 882–890, 2021. 1, 2

[3] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240 × 180 130 db 3 µs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49:2333–2341, 2014. 1

[4] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020. 6

[5] Luis Alejandro Camunas-Mesa, Teresa Serrano-Gotarredona, Sio Hoi Ieng, Ryad Benjamin Benosman, and Bernabe Linares-Barranco. On the use of orientation filters for 3d reconstruction in event-driven stereo vision. *Frontiers in neuroscience*, 8:48, 2014. 2

[6] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 6

[7] Hoonhee Cho, Jaeseok Jeong, and Kuk-Jin Yoon. Eomvs: Event-based omnidirectional multi-view stereo. *IEEE Robotics and Automation Letters*, 6(4):6709–6716, 2021. 2

[8] Hoonhee Cho and Kuk-Jin Yoon. Event-image fusion stereo using cross-modality feature propagation. In *36th AAAI Conference on Artificial Intelligence (AAAI 22)*. Association for the Advancement of Artificial Intelligence, 2022. 1, 2

[9] Hoonhee Cho and Kuk-Jin Yoon. Selection and cross similarity for event-image deep stereo. In *2022 European Conference on Computer Vision (ECCV)*. European Conference On Computer Vision, 2022. 1, 2

[10] Jonghyun Choi, Kuk-Jin Yoon, et al. Learning to super resolve intensity images from events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2768–2776, 2020. 2

[11] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3586–3595, 2020. 3

[12] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 5

[13] Andreas Geiger, P Lenz, and R Urtasun. Are we ready for autonomous driving. In *The KITTI Vision Benchmark Suite, 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 5

[14] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 484–500, 2018. 2

[15] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3268–3277, 2019. 6

[16] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007. 6

[17] Yuhuang Hu, Tobi Delbruck, and Shih-Chii Liu. Learning to exploit multiple vision modalities by using grafted networks. In *European Conference on Computer Vision*, pages 85–101. Springer, 2020. 2

[18] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1312–1321, 2021. 3

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] Jurgen Kogler, Martin Humenberger, and Christoph Sulzbachner. Event-based stereo matching approaches for frameless address event stereo data. In *International Symposium on Visual Computing*, pages 674–685. Springer, 2011. 2

[21] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[22] Hamid Laga, Laurent Valentin Jospin, Farid Boussaïd, and Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2020. 1

[23] Rui Liu, Chengxi Yang, Wenxiu Sun, Xiaogang Wang, and Hongsheng Li. Stereogan: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12757–12766, 2020. 2, 6

[24] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 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*, pages 4040–4048, 2016. 1, 5

[25] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 1, 5

[26] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters*, 7(2):3515–3522, 2022. 2

[27] Mohammad Mostafavi, Lin Wang, and Kuk-Jin Yoon. Learning to reconstruct hdr images from events, with applications

to depth and flow prediction. *International Journal of Computer Vision*, 129(4):900–920, 2021. 2

[28] Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Event-intensity stereo: Estimating depth by the best of both worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4258–4267, 2021. 1, 2, 6, 7

[29] Hyeonseob Nam and Hyo-Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. *Advances in Neural Information Processing Systems*, 31, 2018. 2, 4

[30] Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Stereo depth from events cameras: Concentrate and focus on the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6114–6123, 2022. 1, 2, 6, 7

[31] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015. 2

[32] Vitalijs Osadcuks, Mihails Pudzs, Andrejs Zujevs, Aldis Pecka, and Arturs Ardavs. Clock-based time synchronization for an event-based camera dataset acquisition platform. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4695–4701. IEEE, 2020. 1

[33] Jiahao Pang, Wenxiu Sun, Chengxi Yang, Jimmy Ren, Ruichao Xiao, Jin Zeng, and Liang Lin. Zoom and learn: Generalizing deep stereo matching to novel domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2070–2079, 2018. 2

[34] Federico Paredes-Vallés and Guido CHE de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021. 2

[35] Ewa Piatkowska, Ahmed Belbachir, and Margrit Gelautz. Asynchronous stereo vision for event-driven dynamic stereo sensor using an adaptive cooperative approach. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 45–50, 2013. 2

[36] Ewa Piatkowska, Jurgen Kogler, Nabil Belbachir, and Margrit Gelautz. Improved cooperative stereo matching for dynamic vision sensors with ground truth evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 53–60, 2017. 2

[37] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and D. Scaramuzza. Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time. *International Journal of Computer Vision*, 126:1394–1414, 2017. 2

[38] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 2, 3, 6, 7

[39] Paul Rogister, Ryad Benosman, Sio-Hoi Ieng, Patrick Lichtsteiner, and Tobi Delbruck. Asynchronous event-based binocular stereo matching. *IEEE Transactions on Neural Networks and Learning Systems*, 23(2):347–353, 2011. 2

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[41] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42, 2004. 1

[42] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 156–163, 2020. 2

[43] Xiao Song, Guorun Yang, Xinge Zhu, Hui Zhou, Zhe Wang, and Jianping Shi. Adastereo: a simple and efficient approach for adaptive stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10328–10337, 2021. 2, 3, 4, 5, 8

[44] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *European Conference on Computer Vision*, pages 534–549. Springer, 2020. 2

[45] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019. 3

[46] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022. 2, 6

[47] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 195–204, 2019. 2

[48] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1527–1537, 2019. 1, 2, 7

[49] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 2, 4

[50] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *European Conference on Computer Vision*, pages 155–171. Springer, 2020. 2

[51] Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 608–619, 2021. 2

[52] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, 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*, pages 10081–10090, 2019. 2

[53] Lin Wang, Tae-Kyun Kim, and Kuk-Jin Yoon. Eventsr: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8315–8325, 2020. 2

[54] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[55] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1956–1965, 2020. 6

[56] Jiaolong Xu, Liang Xiao, and Antonio M López. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7:156694–156706, 2019. 3

[57] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2019. 1

[58] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *European Conference on Computer Vision*, pages 420–439. Springer, 2020. 2, 4, 6

[59] Kaixuan Zhang, Kaiwei Che, Jianguo Zhang, Jie Cheng, Ziyang Zhang, Qinghai Guo, and Luziwei Leng. Discrete time convolution for fast event-based stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8676–8686, 2022. 1, 2, 7

[60] Alex Zihao Zhu, Yibo Chen, and Kostas Daniilidis. Real-time time synchronized event-based stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 433–447, 2018. 2

[61] Alex Zihao Zhu, Ziyun Wang, Kaung Khant, and Kostas Daniilidis. Eventgan: Leveraging large scale image datasets for event cameras. In *2021 IEEE International Conference on Computational Photography (ICCP)*, pages 1–11. IEEE, 2021. 3, 6, 7

[62] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[63] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[64] Dongqing Zou, Ping Guo, Qiang Wang, Xiaotao Wang, Guangqi Shao, Feng Shi, Jia Li, and Paul-KJ Park. Context-aware event-driven stereo matching. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1076–1080. IEEE, 2016. 2

[65] Dongqing Zou, Feng Shi, Weiheng Liu, Jia Li, Qiang Wang, Paul-KJ Park, Chang-Woo Shi, Yohan J Roh, and Hyunsurk Eric Ryu. Robust dense depth map estimation from sparse dvs stereos. In *British Mach. Vis. Conf.(BMVC)*, volume 1, 2017. 2