Dual Transfer Learning for Event-based End-task Prediction via Pluggable Event to Image Translation

Lin Wang, Yujeong Chae, and Kuk-Jin Yoon
Visual Intelligence Lab., KAIST, Korea
{wanglin, yujeong, kjyoon}@kaist.ac.kr

Abstract

Event cameras are novel sensors that perceive the per-pixel intensity changes and output asynchronous event streams with high dynamic range and less motion blur. It has been shown that events alone can be used for end-task learning, e.g.*, semantic segmentation, based on encoder-decoder-like networks. However, as events are sparse and mostly reflect edge information, it is difficult to recover original details merely relying on the decoder. Moreover, most methods resort to the pixel-wise loss alone for supervision, which might be insufficient to fully exploit the visual details from sparse events, thus leading to less optimal performance. In this paper, we propose a simple yet flexible two-stream framework named Dual Transfer Learning (DTL) to effectively enhance the performance on the end-tasks without adding extra inference cost. The proposed approach consists of three parts: event to end-task learning (EEL) branch, event to image translation (EIT) branch, and transfer learning (TL) module that simultaneously explores the feature-level affinity information and pixel-level knowledge from the EIT branch to improve the EEL branch. This simple yet novel method leads to strong representation learning from events and is evidenced by the significant performance boost on the end-tasks such as semantic segmentation and depth estimation.*

1. Introduction

Event cameras have recently received much attention in the computer vision and robotics community for their distinctive advantages, *e.g.*, high dynamic range (HDR) and less motion blur [11]. These sensors perceive the intensity changes at each pixel asynchronously and produce event streams encoding time, pixel location, and polarity (sign) of intensity changes. Although events are sparse and mostly respond to the edges in the scene, it has been shown that it is possible to use events alone for learning the end-tasks, *e.g.*, semantic segmentation [1, 14], optical flow and depth estimation [15, 22, 55, 80], via deep neural networks (DNNs). These methods usually follow the encoder-decoder-like network structures, *e.g.*, [1, 14, 22], as shown in Fig. 1(a), and are trained in a fully supervised manner. However, as events mostly reflect edge information, unlike the canonical images, it is difficult to recover the original structural details from events merely relying on the decoder (see Fig. 2(b)). Importantly, learning from sparse events with the pixel-wise loss (*e.g.*, cross-entropy loss) alone for supervision often fails to fully exploit visual details from events, thus leading to less optimal performance.

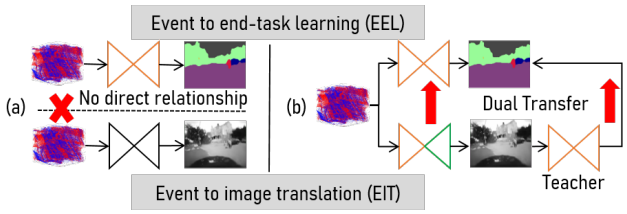


Figure 1: (a) There is no direct relation between EIT and EEL in the prior-arts. (b) The proposed DTL framework by using EIT branch as a pluggable unit and transferring both feature-level and prediction-level information to enhance the performance of EEL.

The other line of research has shown the possibility of generating images from events [24, 37, 48, 54, 58, 61, 66]. The generated images have been successfully applied to the end-tasks, *e.g.*, object recognition [48]; however, there exist two crucial problems. First, using these images as the intermediate representations of events leads to considerable inference latency. Second, there is no direct connection regarding the optimization process between two tasks. Indeed, the feature representations learned from image generation contain more structural details (see Fig. 2(c)), which can be a good guide for learning end-tasks. However, these crucial clues have been rarely considered and explored to date. Moreover, some event cameras provide active pixel sensor (APS) frames, which contain very crucial visual information; nonetheless, the potential has been scarcely explored to assist learning end-tasks from events.

Therefore, in this paper, we design a concise yet flexible framework to alleviate the above dilemmas. Motivated by recent attempts for event-to-image translation [42, 48, 58, 61], transfer learning [31, 81], and multi-task learning

[56], we propose a novel Dual Transfer Learning (DTL) paradigm to efficiently enhance the performance of the end-task learning (see Fig. 3). Such a learning method is unified in a two-stream framework, which consists of three components: Event to End-task Learning (EEL) branch, Event to Image Translation (EIT) branch and Transfer Learning (TL) module, as shown in Fig. 1(b). Specifically, we integrate the idea of image translation to the process of end-task learning, thus formulating the EIT branch. The EEL branch is then significantly enhanced by the TL module, which exploits to transfer the feature-level and prediction-level information from EIT branch. In particular, a novel affinity graph transfer loss is proposed to maximize the feature-level instance similarity along the spatial locations between EEL and EIT branches. The prediction-level information is transferred from the EIT branch to the EEL branch using the APS and translated images based on a teacher network trained using the canonical images on the end-task. Moreover, we propose to share the same feature encoder between EEL and EIT branches and optimize them in an end-to-end manner. We minimize the prediction gap between the APS and translated images based on the teacher network and subtly leverage the supervision signal of the EEL branch to enforce semantic consistency for the EIT branch, which surprisingly helps to recover more semantic details for image translation. *Once training is done, the EIT branch and TL module can be freely removed, adding no extra inference cost.*

We conduct extensive experiments on two end-tasks, semantic segmentation (Sec. 4.1) and depth estimation (Sec. 4.2). The results show that this simple yet novel method brings significant performance gains for both tasks. As a potential, our method can also learn the end-tasks via the teacher network *without* using ground truth labels. Although the EIT branch is regarded as an *auxiliary* task, the results demonstrate that our DTL framework contributes to recover better semantic details for image translation.

In summary, our contributions are three folds. (I) We propose a novel yet simple DTL framework for the end-task learning. (II) We propose a TL module where we transfer both feature-level and prediction-level information to the end-tasks. (III) We conduct extensive experiments on two typical end-tasks, showing that DTL significantly improves the performances while adding no extra inference cost. We also demonstrate that DTL recovers better semantic details for the EIT branch. Our project code is available at `https://github.com/addisonwang2013/DTL`.

2. Related Works

DNNs for event-based vision. DNNs with event data was first explored in the classification [39] and robot control [35]. [32] then trained a DNN for steering angle prediction on the DDD17 dataset [4]. This dataset has been utilized in [1, 14, 59] to perform semantic segmentation using pseudo labels obtained from the APS frames. Moreover, DNNs have been applied to high-level tasks, such as object detection and tracking [6, 8, 24, 33], human pose estimation [5, 64, 69], motion estimation [26, 34, 52, 65], object recognition [3, 14, 48] on N-Caltech [40] and other benchmark datasets [3, 28, 51].

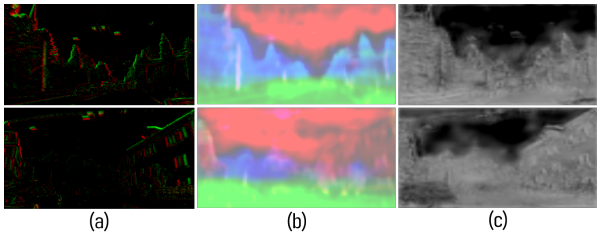


Figure 2: Visualization of features from EEL and EIT decoders of the same input. (a) Events, (b) EEL features, (c) EIT features.

Meanwhile, another line of research focuses on the low-level vision tasks, such as optical flow estimation [12, 15, 53, 80], depth estimation [16, 37, 55, 80] on the MVSEC dataset [78]. In addition, [48, 50, 54, 61] attempted to generate video from events using camera simulator [38, 45], and [36, 58, 60] tried to generate high-resolution images. In contrast, [14] proposed to generate events from video frames. Some other works explored the potential of events for image deblurring [20, 25, 60], HDR imaging [19, 73], and event denoising [2]. For more details about event-based vision, refer to a survey [11]. Differently, we propose a DTL framework to enhance the performance for the end-task learning by exploring the knowledge from the EIT branch sharing the same encoder. We regard the EEL branch as main task and EIT branch as auxiliary one. As a potential, the EIT is also improved by the DTL framework.

Transfer Learning (TL). TL aims to improve learning in a new task through the transfer of knowledge from a related task that has already been learned [81]. Among the techniques, knowledge transfer (KT) is a typical approach for learning model with softmax labels or feature information of a learned model [23, 49, 62]. Most KT methods focus on the image data and transfer knowledge using logits [7, 68, 70, 74] or features [21, 27, 43, 49, 63, 72]. Recently, some approaches have been proposed to transfer knowledge across different paired modality data with common labels [17, 18, 24, 29, 41, 67, 71, 76]. For more details of TL, refer to recent surveys, *e.g.*, [62, 81]. Differently, we transfer knowledge across the EIT branch and the EEL branch. We do not explore multiple input modalities with multiple task networks. Instead, our input is homogeneous and tasks to be learned are not the same. Moreover, the networks share the same encoder, and the knowledge is transferred in two aspects: feature-level transfer via affinity graph learning and prediction-level transfer via a teacher network.

Multi-task learning. It aims to learn multiple tasks, *e.g.*, object recognition, detection and semantic segmentation,

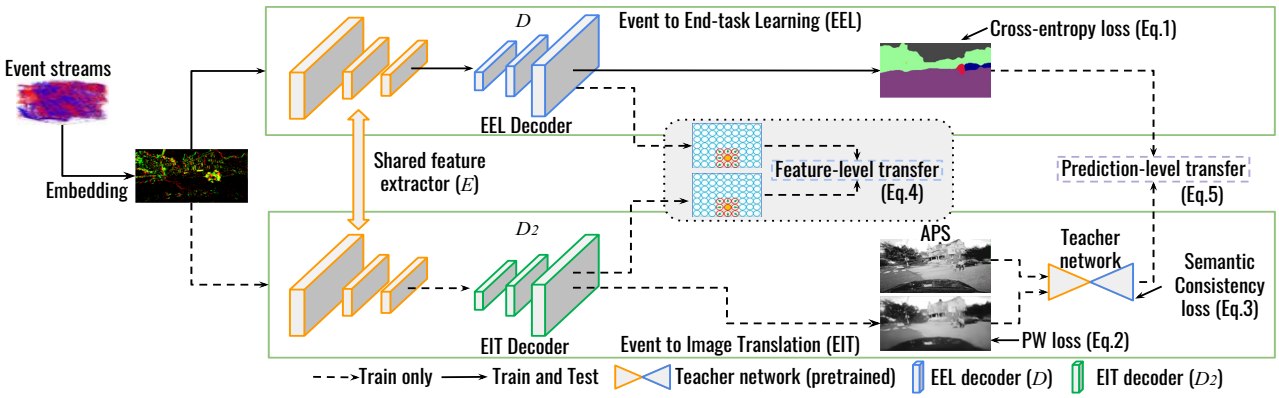


Figure 3: Overview of the proposed DTL framework, which consists of three parts: Event to End-task Learning (EEL) branch, Event to Image Translation (EIT) branch and Transfer Learning (TL) module. The dash line indicates 'for training only'.

jointly by leveraging the domain-specific information contained in the training signals of relevant tasks [57, 67, 75]. For more details, refer to a recent survey [56]. These methods usually treat the multiple tasks equally in both training and inference. However, different from these methods for joint task learning in the same modality, we learn from cross-modalities and regard the EEL branch as the main and EIT branch as the auxiliary one. The EIT branch and TL module can be flexibly removed during inference, adding no extra computation cost.

3. The Proposed Approach

Event Representation. As DNNs are designed for image-/tensor-like inputs, we first describe the way of event embedding. An event e is interpreted as a tuple $(\mathbf{u}, t, p)$, where $\mathbf{u} = (x, y)$ is the pixel coordinate, t is the timestamp, and p is the polarity indicating the sign of brightness change. An event occurs whenever a change in log-scale intensity exceeds a threshold C. A natural choice is to encode events in a spatial-temporal 3D volume to a voxel grid [48, 79, 80] or event frame [15, 46] or multi-channel image [30, 58, 61]. In this work, we represent events to multi-channel images as the inputs to the DNNs, as done in [30, 58, 61]. More details about event representation are in the suppl. material.

3.1. Overview

For event cameras, *e.g.*, DAVIS346C with APS frames, assume that we are given the dataset $\mathcal{X} = \{e_i, x_{aps_i}, y_i\}$, where e_i is i-th stacked multi-channel event image and x_{aps_i} is the corresponding i-th APS image with its ground truth label y_i. Our goal is to learn an effective end-task model on the end-tasks, *e.g.*, semantic segmentation, from the events. Existing methods [1, 14] rely on the encoder-decoder network structures and are trained using the ground truth (GT) labels for supervision via, *e.g.*, cross-entropy loss. However, as events are sparse and mostly reflect the edges of the scene, it is difficult to recover original details merely relying on the decoder, as shown in Fig. 2(b).

To address the dilemmas, we propose a novel yet effective framework, called dual transfer learning (DTL), to effectively learn the end-tasks. As shown in Fig. 3, the DTL framework consists of three components: (a) Event to End-task Learning (EEL) branch; (b) Event to Image Translation (EIT) branch, and Transfer Learning (TL) module. The TL module transfers knowledge from the EIT branch to learn better representation of events for the EEL branch, without adding extra computation cost in the inference time. We now describe these components in detail.

3.2. Dual Transfer Learning

3.2.1 Event to End-task Learning (EEL)

For the end-tasks, *e.g.*, semantic segmentation, we simply adopt an encoder(E)-decoder(D) network [10]. The whole process of learning this branch is called Event to End-task Learning (EEL), as shown in Fig. 3. The EEL branch generates an output of prediction, *e.g.*, label map, from an embedded event image in a dimension of W×H×C, where W, H and C are the width, height and number of channels. This setting is similar to the methods for semantic segmentation [1, 14], in which a conventional multi-class cross-entropy (CE) loss is used for supervision, which is defined as:

$$\mathcal{L}_{CE} = \frac{1}{N} \sum_{i=1}^{N} -y_i \log(p_i) \tag{1}$$

where N is the total pixel numbers, p_i and y_i refer to the predicted probability and GT label for pixel i. For the end-tasks, *e.g.*, depth estimation, the loss for supervision can be flexibly changed to other pixel-wise loss, such as L_1 loss.

However, we notice that using the supervision loss only is insufficient to fully exploit the visual information from events. Furthermore, it is difficult to recover original details only relying on the decoder, as shown in Fig. 2(b). The reason is that events respond predominantly to edges of the scenes, making the event data intrinsically sparse. This renders dense pixel-wise predictions from events challenging, especially in low contrast change or less motion regions.

To this end, we draw attention from recent attempts for

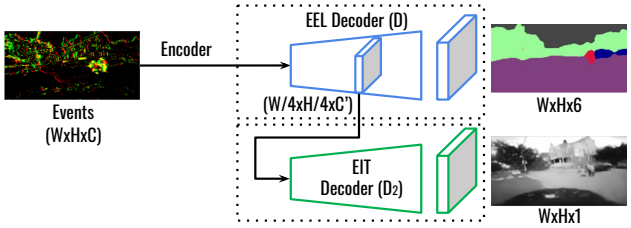


Figure 4: Designing the EIT decoder based on EEL branch.

events to image translation (EIT) [24, 48, 58]. Our motivations are two-folds. Firstly, given the same event input, we find that the feature representations of the EIT decoder contain more complete structural information of scenes, as shown in Fig. 2(c). Secondly, the APS frames synchronized with event sensor and the translated images acting as intermediate representation of events can be fully leveraged for guiding the EEL branch. To this end, we design an EIT branch and further exploit to transfer both feature-level and prediction-level knowledge to the EEL branch. We will discuss the details in the following Sec. 3.2.2.

3.2.2 Event to Image Translation (EIT)

Under the similar network structure, the feature maps of EIT contain fine-grained visual structural information of scenes, as shown Fig. 2(c). Although these structural information does not convey the object class information, they can be effectively categorized and optimized between pixel to pixel or region to region. That is, these categorized information indeed delivers crucial semantic knowledge, which can benefit the end-tasks. These details can be modeled by the correlation or relations between the internal pixels.

For the EIT branch, as the input is the same as the EEL branch, the encoded latent spaces from both branches are similar. We thus propose to share the same encoder E for EEL and EIT branches to extract visual features, as shown in Fig. 4. As the EEL decoder D is not effective enough to reconstruct the fine-grained structure information of image, we design a decoder D_2 to generate high-quality results while reducing the computation. The detailed structure of D_2 is depicted in Fig. 5. D_2 is based on the penultimate layer (with a dimension of W/4×H/4×C') of EEL decoder D, which is further extended by designing ResBlocks connected by the deconvolution layers, followed by one 3x3 convolution (conv.) layer and a Tanh function. In particular, the ResBlock consists of a residual connection that takes one 3x3 conv. layer, followed by a ReLU function and one 3x3 conv. layer, helping to enlarge the receptive field. Interestingly, we find that adding Tanh function is crucial for image reconstruction. Based on the shared encoder E and decoder D_2, we reconstruct images from events under the supervision of the APS frames using a pixel-wise loss (*e.g.*, L_1 loss), which can be formulated as follows:

$$\mathcal{L}_{EIT} = \mathbb{E}_{e_i \sim \mathcal{X}}[||x_{aps_i} - D_2(E(e_i))||_1]. \quad (2)$$

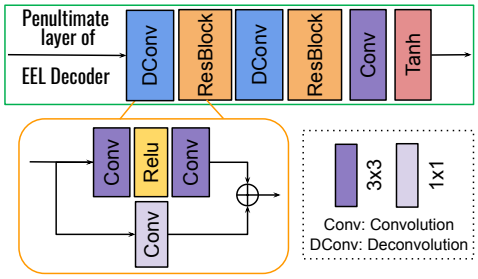


Figure 5: The proposed EIT decoder network structure.

To better preserve the semantic information, we exploit a novel semantic consistency (SC) loss based on a teacher network T pretrained on the end-task with canonical images, as shown in Fig. 3. The proposed SC loss for EIT has three advantages. Firstly, the generated image $D_2(E(e_i))$ becomes the optimal input of T. Secondly, the predictions from both the APS frames and generated images can provide auxiliary supervision for the EEL branch. The EIT branch can be removed after training, adding no extra inference cost. The proposed SC loss is formulated as follows:

$$\mathcal{L}_{SC} = \mathbb{E}_{e_i \sim \mathcal{X}} KL[T(D_2(E(e_i)))||T(x_{aps_i})] \quad (3)$$

where $KL(\cdot||\cdot)$ is the Kullback-Leibler divergence between two distributions.

3.2.3 Transfer Learning (TL) Module

Feature-level Transfer. As the feature representations of the decoder D_2 for the EIT branch (see Fig. 2(c)) deliver fine-grained visual structural information of scenes, we leverage these visual knowledge to guide the feature representation of the decoder D of the EEL branch. To this end, we propose a novel approach to transfer the instance-level similarity along the spatial locations between EIT branch and EEL branch based on affinity graphs, as formulated in Eq. 4. The node represents a spatial location of an instance (*e.g.*, car), and the edge connected between two nodes represents the similarity of pixels. For events, if we denote the connection range (neighborhood size) as σ, then nearby events within σ (9 nodes in Fig. 3) are considered for computing affinity contiguity. It is possible to adjust each node's granularity to control the size of the affinity graph; however, as events are sparse, we do not consider this factor. In such a way, we can aggregate top-σ nodes according to the spatial distances and represent the affinity feature of a certain node. For a feature map $F \sim \mathbb{R}^{C \times H \times W}$ ($H \times W$ is the spatial resolution and C is the number of channels), the affinity graph contains nodes with $H \times W \times \sigma$ connections. We denote A_{ab}^{EIT} and A_{ab}^{EEL} are the affinity graph between the a-th node and the b-th node obtained from the EIT and EEL branch, respectively, which is formulated as:

$$\mathcal{L}_{FL} = \frac{1}{H \times W \times \sigma} \sum_{a \sim R} \sum_{b \sim \sigma} ||A_{ab}^{EIT} - A_{ab}^{EEL}||_2^2 \quad (4)$$

Table 1: Segmentation performance with different event representations and APS frames on the test data [1], measured by Acc. (Accuracy) and MIoU (Mean Intersection over Union). The models are trained on time intervals of 50ms but tested with 50ms, 10ms and 250ms.

Method	Event representation	MIoU [50ms]	MIoU [10ms]	MIoU [250ms]
EvSegNet [1]	6-channel [1]	54.81	45.85	47.56
Vid2E [14]	EST [15]	45.48	30.70	40.66
Ours-Deeplabv3 (Baseline)	Multi-channel	50.92	41.61	43.87
Ours-Deeplabv3 (DTL)	Multi-channel	**58.80** (+7.88)	**50.01** (+8.40)	**52.96** (+9.09)

where $R = \{1, 2, \cdots, H \times W\}$ indicates all the nodes in the graph. The similarity between two nodes is calculated from the aggregated features F_a and F_b as: $A_{ab} = \frac{F_a^\intercal F_b}{||F_a||_2||F_b||_2}$, where $F_a^\intercal$ is the transposed feature vector of F_a. More detailed formulation of the proposed feature-level (FL) transfer loss is provided in the suppl. material.

Prediction-level Transfer. In addition to transferring the structural information of the feature representations from the EIT decoder D_2, we observe that it is potential to leverage the paired APS frames to enhance EEL branch. In particular, inspired by the recent attempts for cross-modal learning [17, 76], we aim to transfer the knowledge from the teacher network T using the APS frames to the EEL network. We view the segmentation problem as a collection of separate pixel labeling problems, and directly strive to align the class probability of each pixel produced by the EEL network with that by the teacher network. We use the class probabilities produced from T as soft targets for training the EEL network $E(D(\cdot))$. The prediction-level transfer loss is formulated as:

$$\mathcal{L}_{PL} = \frac{1}{H \times W} \sum_{k \in \Omega} KL[E(D(e_i^k))||T(x_{aps_i}^k)] \quad (5)$$

where $E(D(e_i^k))$ represents the class probabilities of k-th pixel of i-th event image, $T(x_{aps_i}^k)$ represents the class probabilities of the k-th pixel of i-th APS image from the teacher T, and $\Omega = \{1, 2, ..., W \times H\}$ denotes all the pixels.

3.2.4 Optimization

The overall objective consists of the supervision loss $\mathcal{L}_{CE}$ in Eq. 1 for EEL branch, together with the EIT loss $\mathcal{L}_{EIT}$ in Eq. 2 and the SC loss $\mathcal{L}_{SC}$ in Eq. 3. Moreover, it includes the loss terms in the TL module, namely, the feature-level transfer loss $\mathcal{L}_{FL}$ in Eq. 4 and prediction-level transfer loss $\mathcal{L}_{PL}$ in Eq. 5. The overall objective function is:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{EIT} + \lambda_2 \mathcal{L}_{SC} + \lambda_3 \mathcal{L}_{FL} + \lambda_4 \mathcal{L}_{PL} \quad (6)$$

where λ_1, λ_2, λ_3 and λ_4 are the hyper-parameters. We minimize the overall objective with respect to both EEL and EIT branches using dynamic gradient descent strategy.

4. Experiments and Evaluation

4.1. Event-based Semantic Segmentation

Semantic segmentation is the end-task aiming to assign a semantic label, *e.g.*, road, car, in a given scene to each pixel.

Datasets. We use the publicly available driving scene dataset DDD17 [4], which includes both events and APS frames recorded by a DAVIS346 event camera. In [1], 19,840 APS frames are utilized to generate pseudo annotations (6 classes) based on a pretrained network for events (15,950 for training and 3,890 for test). As the events in the DDD17 dataset are very sparse and noisy, we show more results on the driving sequences in the MVSEC dataset [78], collected for the 3D scene perception purpose.

Implementation details. We use DeepLabv3 [10] as the semantic segmentation network. The hyper-parameters λ_1, λ_2, λ_3 and λ_4 are set as 1, 1, 0.1 and 1, respectively. In the training, we set the learning rate as $1e-3$ and use the stochastic gradient descent (SGD) optimizer with weight decay rate of $5e-6$ to avoid overfitting. As the common classification accuracy does not well fit for semantic segmentation, we use the following metric to evaluate the performance, as done in the literature [9, 10]. The *intersection of union* (IoU) score is calculated as the ratio of intersection and union between the ground-truth mask and the predicted segmentation mask for each class. We use the *mean IoU* (MIoU) to measure the effectiveness of segmentation.

4.1.1 Evaluation on DDD17 dataset

Comparison. We first present the experimental results on the DDD17 dataset [1]. We evaluate our method on the test set and vary the window size of events between 10, 50, and 250ms, as done in [1]. The quantitative and qualitative results are shown in Table 1 and Fig. 6. We compare our method with two SoTA methods, EvSegNet [1] and Vid2E [14] that uses synthetic version of DDD17 data. Quantitatively, it turns out that the proposed DTL framework significantly improves the segmentation results on events than the baseline (with only CE loss) by around 8% increase of MIoU. It also surpasses the existing methods by around 4% increase of MIoU with a multi-channel event representation.

Meanwhile, on the time interval of 10ms and 250ms, our DTL framework also shows a significant increase of MIoU by around 8.4% and 9.1% than those of the baseline, respectively. The visual results in Fig. 6 further verify the effectiveness of the proposed DTL framework. Overall, the segmentation results on events are comparable to those of the APS frames, and some are even better *e.g.*, the 1st and 2nd rows. Meanwhile, our method generates convincing intensity images from EIT branch (5th column), The results indicate that, although events only reflect the edge informa-

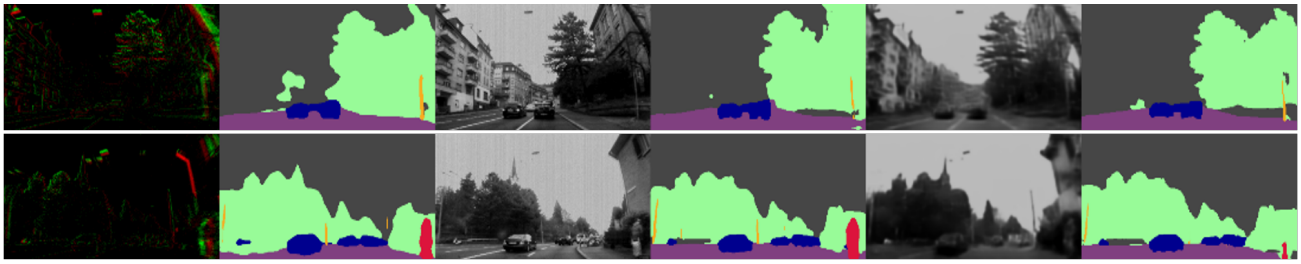


Figure 6: Qualitative results on DDD17 test sequence provided by [1]. (a) Events, (b) Segmentation results on events, (c) APS frames, (d) Segmentation results on APS frames, (e) Generated intensity images from events, (f) Pseudo GT labels.

Table 2: Segmentation performance of our method and the baseline on the test data [78], measured by MIoU. The baseline is trained using the pseudo labels made by the APS frames.

Method	Event representation	MIoU
Baseline-Deeplabv3	Multi-channel	50.53
Ours-Deeplabv3	Multi-channel	**60.82** (+ 10.29)

tion, our method successfully explores the feature-level and prediction-level knowledge to facilitate the end-task learning. The simple yet flexible approach brings a significant performance boost on the end-task learning.

High dynamic range (HDR). HDR is one distinct advantage of an event camera. We show the segmentation network shows promising performance in the extreme condition. Figure 4 in the suppl. material shows the qualitative results. The APS frames are over-exposed, and the network fails to segment the urban scenes; however, the events capture the scene details, and the EEL network shows more convincing segmentation results.

Segmentation without using GT labels. With the EIT branch empowered by the teacher model, we show that our DTL framework can learn to segment events without using the semantic labels. The quantitative and qualitative results are in the suppl. material. Numerically, even without using the ground truth labels, our method achieves 56.52% MIoU, which significantly enhances the semantic segmentation performance by 6.60% MIoU than the baseline. Compared with the SoTA methods [1, 13], our method still surpasses them by around 2% MIoU.

4.1.2 Evaluation on MVSEC dataset

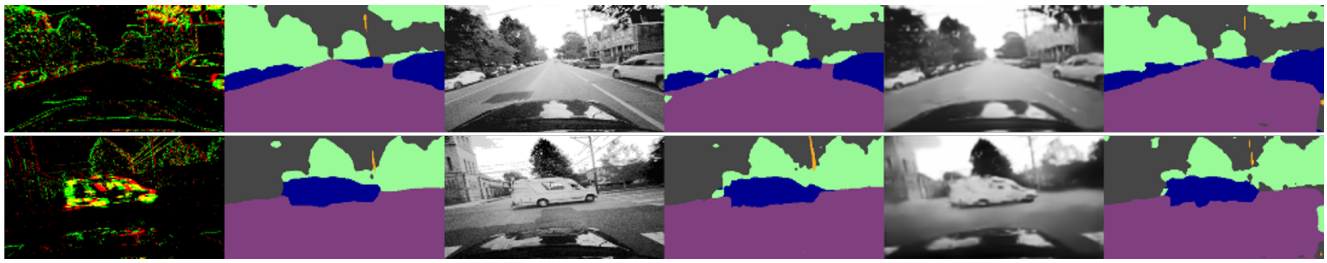
To further validate the effectiveness of the proposed DTL framework, we show more results on the MVSEC dataset [78]. As there are no segmentation labels in this dataset, to numerically evaluate our method, we utilize the APS frames to generate pseudo labels based on a network [10], similar to [1], as our comparison baseline. Due to the poor quality of APS frames in the outdoor_day1 sequence, we mainly use outdoor_day2 sequence and divide the data into training (around 10K paired event images and APS frames) and test (378 paired event images and APS frames) sets based on the way of splitting DDD17 dataset in [1]. For the training data, we remove the redundant sequences, such as vehicles stopping in the traffic lights, etc. We also use the night driving sequences to show the advantage of events on HDR.

The qualitative and quantitative results are shown in Fig. 7 and Table 2. In Fig. 7, we mainly show the results in the general condition. The experimental results of HDR scenes are provided in suppl. material. Using a multi-channel event representation in Table 2, the proposed DTL framework significantly surpasses the baseline by a noticeable margin with around 10.3% increase of MIoU. The results indicate a significant performance boost for semantic segmentation. The effectiveness can also be verified from visual results in Fig. 7. As can be seen, the semantic segmentation results (2nd column) are fairly convincing compared with the results on APS frames (4th column) and the pseudo GT labels (6th column). Meanwhile, our method also generates very realistic intensity images (5th column) from the EIT branch. The results on both semantic segmentation and image translation show that our the proposed DTL framework successfully exploit the knowledge from one branch to enhance the performance of the other.

4.2. Monocular Dense Depth Estimation

Depth estimation is the end-task of predicting the depth of scene at each pixel in the image plane. Previous works for event-based depth estimation have most focused on sparse or semi-dense depth estimation [44, 46, 47, 77]. Recently, DNNs have been applied to stereo events to generate dense depth predictions [55] and to estimate monocular semi-dense depth [80]. Some other works have focused on the dense depth estimation with only events [22] or with additional inputs [16]. We show that the proposed DTL framework is capable of predicting monocular dense depth from sparse event data. To evaluate the scale-invariant depth, we use the absolute relative error (Abs. Rel.), logarithmic mean squared error (RMSELog), scale invariant logarithmic error (SILog) and accuracy (Acc.).

We present quantitative and qualitative results and compare with the baseline settings and prior methods [22, 80] with sparse event data as inputs on the MVSEC dataset [78]. We use outdoor_day2 sequence of the MVSEC dataset where we select around 10K embedded event image and



(a) (b) (c) (d) (e) (f)

Figure 7: Qualitative results of semantic segmentation and image translation on the MVSEC dataset. (a) Events, (b) Segmentation results on events, (c) APS frames, (d) Segmentation results on APS frames, (e) Generated intensity images from events, (f) Pseudo GT labels.

Table 3: Quantitative evaluation of monocular dense depth estimation on the MVSEC dataset.

Method	**Dataset**	**Abs. Rel.** (↓)	**RMSELog** (↓)	**SILog** (↓)	$\sigma < 1.25$ (↑)	$\sigma < 1.25^2$ (↑)	$\sigma < 1.25^3$ (↑)
[80]	Outdoor_day1	0.36	0.41	0.16	0.46	0.73	0.88
[22]		0.45	0.63	0.25	0.47	0.71	0.82
Translated images		0.52	0.69	0.31	0.33	0.52	0.71
Baseline		0.33	0.39	0.16	0.63	0.80	0.89
Ours (DTL)		**0.29**	**0.34**	**0.13**	**0.71**	**0.88**	**0.96**
[80]	Outdoor_night1	0.37	0.42	0.15	0.45	0.71	0.86
[22]		0.77	0.64	0.35	0.33	0.58	0.73
Translated images		0.46	0.83	0.68	0.26	0.49	0.69
Baseline		0.36	0.40	0.14	0.57	0.77	0.88
Ours (DTL)		**0.30**	**0.35**	**0.12**	**0.69**	**0.88**	**0.95**

APS image pairs with their synchronized depth GT images to train the our DTL framework, similar to [80]. We then utilize the outdoor_day1 sequence (normal driving condition) and night driving sequences as the test sets. More details about dataset preparation and implementation (*e.g.*, network structure, loss functions) are in the suppl. material.

Table 3 shows the quantitative results, which are supported by the qualitative results in Fig. 8. On the outdoor_day1 sequence, our method achieves around 10% Abs. Rel. drop than the baseline and the compared methods. The effectiveness on outdoor_day1 sequence can also be verified from Fig. 8 (1st row). Compared with the GT depth, our method predicts depth with clear edges and better preserves the shapes and structures of objects, such as buildings, trees, cars, etc. Meanwhile, our method is also capable of translating events to high-quality intensity images, where we can see the translated images are close to the APS frames.

HDR depth. Our method shows apparent advantages on the HDR scene. As shown in Fig. 8 (2nd and 3rd rows), when the APS frames fail to predict the correct depth information (6th column), events show promising depth estimation results (4th columns). Moreover, our method generates realistic intensity images (3rd column) and shows better depth information (5th column) than those of APS frames. In particular, when the APS frames are almost invisible, our method generates convincing HDR images that better preserve the structures of objects, such as buildings, trees, cars, etc. The effectiveness can also be numerically verified in Table 3. Our method achieves around 17% performance boost (*e.g.*, Abs. Rel.) than those of SoTA methods [22, 80], the baseline and generated intensity images.

5. Ablation Study and Analyses

Loss functions. We first study the effectiveness of adding and removing the loss terms in Eq. 6. For convenience, we mainly focus on semantic segmentation on the DDD17 dataset. The ablation results are shown in Table 4. In general, without TL module, the EIT branch slightly improves the EEL branch. However, with feature-level transfer loss in the TL module, the performance of segmentation is significantly enhanced by around 5.58% increase of MIoU. When the predication-level transfer loss is added, EEL performance is further enhanced to MIoU of 58.80%. From the ablation study, it clearly shows that our DTL framework is a successful approach for benefiting the end-task learning.

The effectiveness of TL module for EIT branch. Although the EIT branch is regarded as an *auxiliary* task in the proposed DTL framework, we show that it also benefits the EIT learning. We qualitatively compare the quality of translated images with and without using the DTL framework. Fig. 9 shows the visual results. In contrast to the generated images without DTL (2nd column), the results with DTL are shown to have more complete semantic information and better structural details, as shown in the cropped patches in the 3rd column. Interestingly, better structural details, *e.g.*, cars, buildings and trees, are restored. The experimental results show that our method works effectively on sparse events and are shown successful not only for the end-tasks but also for the image translation.

Event representation vs. EEL performance. We now study how event representation impacts the performance on end-task learning under the DTL framework. We leverage

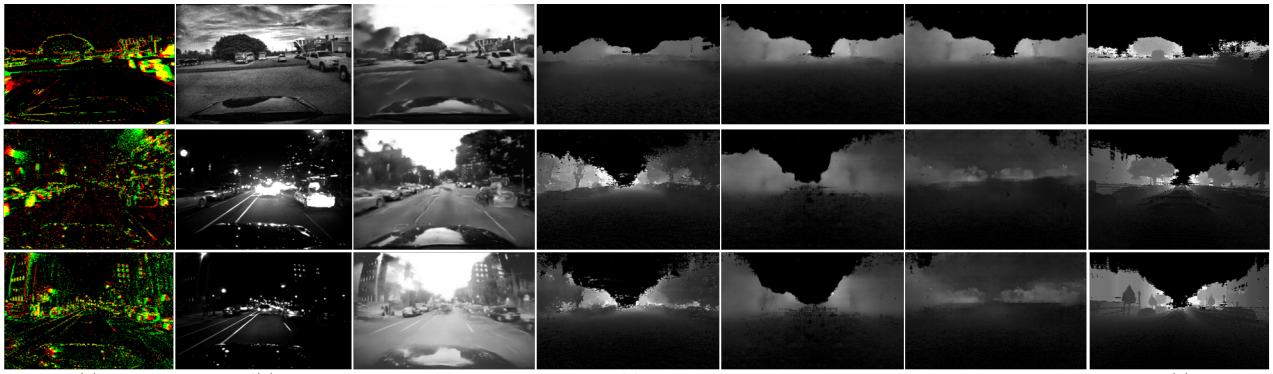


Figure 8: Qualitative results for monocular dense depth estimation. (a) Events, (b) Dark APS frames, (c) Generated intensity images, (d) Predicted depth on events, (e) Predicted depth on the generated intensity images, (f) Predicted depth on APS frames, (g) GT depth.

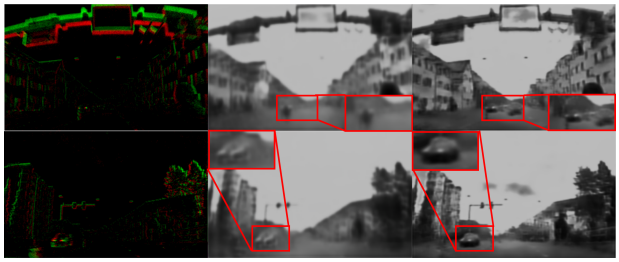


Figure 9: Impact of DTL on image translation. (a) Events, (b) Translated images without DTL, (c) Translated images with DTL.

Table 4: Ablation study results of the proposed DTL framework based on DDD17 dataset.

Module	MIoU [50ms]
CE	50.92
CE + EIT	53.87 (+2.95)
CE + EIT + CS	54.66 (+3.74)
CE + EIT + TL (FL)	56.50 (+5.58)
CE + EIT + TL (FL + PL)	**58.80** (+7.88)

Table 5: The impact of event representations on the semantic segmentation performance on DDD17 dataset.

Method	Event Rep.	MIoU [50ms]
Ours-Deeplabv3	Voxel grid [79]	56.30
Ours-Deeplabv3	6-channel [1]	57.68
Ours-Deeplabv3	Multi-channel	**58.80**

several existing event representation methods, *e.g.*, voxel grid [79] and 6-channel [1] and the multi-channel event embedding methods [58, 61] used in the paper. For convenience, we compare the efficacy of these methods on semantic segmentation task using the DDD17 dataset. The numerical results are shown in Table 5. In general, event embedding methods have a considerable influence on semantic segmentation performance. Overall, the multi-channel representation is demonstrated to show sightly better segmentation performance than the other two methods.

Number of events vs. overall performance. The number of events used for event representation also impacts the performance on the end-task learning and image translation. We thus conduct an analysis on how the number of events used for event representation affects the end-task, *e.g.*, dense depth estimation. Fig. 10 shows the visual comparisons on the outdoor_night2 sequence of the MVSEC dataset, in which we highlight the HDR capability. In particular, the results of embedding 35K events of the bottom row show better depth estimation (as shown in the red box) and intensity image translation (as shown in the yellow box) results than those of 25K and 20K events, respectively.

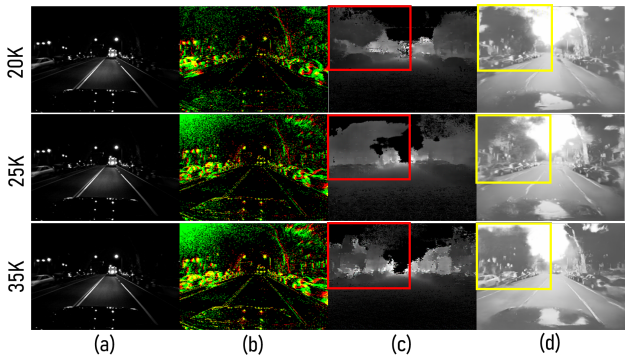


Figure 10: Impact of number of events (top: 20K; middle: 25K; bottom: 35K) on our DTL framework. (a) APS frames, (b) Events, (c) Predicted depth on events, (d) Generated intensity images.

6. Conclusion and Future work

In this paper, we presented a simple yet novel two-stream framework, named DTL for promoting end-task learning, with no extra inference cost. The DTL framework consists of three components: the EEL, the EIT and TL module, which enriches the feature-level and prediction-level knowledge from EIT to improve the EEL. The simple method leads to strong representations of events and is evidenced by the promising performance on two typical tasks. As the DTL framework is a general approach, we plan to apply it to other modality data, such as depth and thermal data.

Acknowledgement. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2018R1A2B3008640).

References

[1] Inigo Alonso and Ana C Murillo. Ev-segnet: semantic segmentation for event-based cameras. In *CVPRW*, pages 0–0, 2019. 1, 2, 3, 5, 6, 8

[2] R Baldwin, Mohammed Almatrafi, Vijayan Asari, and Keigo Hirakawa. Event probability mask (epm) and event denoising convolutional neural network (edncnn) for neuromorphic cameras. In *CVPR*, pages 1701–1710, 2020. 2

[3] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based object classification for neuromorphic vision sensing. In *ICCV*, pages 491–501, 2019. 2

[4] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd17: End-to-end davis driving dataset. *arXiv preprint arXiv:1711.01458*, 2017. 2, 5

[5] Enrico Calabrese, Gemma Taverni, Christopher Awai Easthope, Sophie Skriabine, Federico Corradi, Luca Longinotti, Kynan Eng, and Tobi Delbruck. Dhp19: Dynamic vision sensor 3d human pose dataset. In *CVPRW*, 2019. 2

[6] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. Asynchronous convolutional networks for object detection in neuromorphic cameras. In *CVPRW*, pages 0–0, 2019. 2

[7] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. *arXiv preprint*, 2019. 2

[8] Haosheng Chen, David Suter, Qiangqiang Wu, and Hanzi Wang. End-to-end learning of object motion estimation from retinal events for event-based object tracking. *AAAI*, 34(07):10534–10541, Apr 2020. 2

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*. 5

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *ECCV*, 2018. 3, 5, 6

[11] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, Kostas Daniilidis, and D. Scaramuzza. Event-based vision: A survey. *TPAMI*, PP, 2020. 1, 2

[12] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In *CVPR*, pages 12280–12289, 2019. 2

[13] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Bringing modern computer vision closer to event cameras. *CVPR*, 2020. 6

[14] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *CVPR*, pages 3586–3595, 2020. 1, 2, 3, 5

[15] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *ICCV*, pages 5633–5643, 2019. 1, 2, 3, 5

[16] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo Carrio, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE RA-L*, 2021. 2, 6

[17] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *CVPR*, pages 2827–2836, 2016. 2, 5

[18] Frank Hafner, Amran Bhuiyan, Julian FP Kooij, and Eric Granger. A cross-modal distillation network for person re-identification in rgb-depth. *arXiv preprint arXiv:1810.11641*, 2018. 2

[19] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic camera guided high dynamic range imaging. In *CVPR*, pages 1730–1739, 2020. 2

[20] Chen Haoyu, Teng Minggui, Shi Boxin, Wang YIzhou, and Huang Tiejun. Learning to deblur and generate high frame rate video with an event camera. *arXiv preprint arXiv:2003.00847*, 2020. 2

[21] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, pages 1921–1930, 2019. 2

[22] Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. *International Conference on 3D Vision*, 2020. 1, 6, 7

[23] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[24] Yuhuang Hu, Tobi Delbruck, and Shih-Chii Liu. Learning to exploit multiple vision modalities by using grafted networks. In *ECCV*, pages 85–101. Springer, 2020. 1, 2, 4

[25] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *CVPR*, pages 3320–3329, 2020. 2

[26] Daniel R Kepple, Daewon Lee, Colin Prepsius, Volkan Isler, and Il Memming. Jointly learning visual motion and confidence from local patches in event cameras. *ECCV*, 2020. 2

[27] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *NeurIPS*, pages 2760–2769, 2018. 2

[28] Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*. 2

[29] Kang Li, Lequan Yu, Shujun Wang, and Pheng-Ann Heng. Towards cross-modality medical image segmentation with online mutual knowledge distillation. In *AAAI*, volume 34, pages 775–783, 2020. 2

[30] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *ECCV*, volume 3, 2020. 3

[31] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *CVPR*, pages 2604–2613, 2019. 1

[32] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *CVPR*, pages 5419–5427, 2018. 2

[33] Nico Messikommer, Daniel Gehrig, Antonio Loquercio, and Davide Scaramuzza. Event-based asynchronous sparse con-

volutional networks. *arXiv preprint*. 2

[34] Anton Mitrokhin, Zhiyuan Hua, Cornelia Fermuller, and Yiannis Aloimonos. Learning visual motion segmentation using event surfaces. In *CVPR*, pages 14414–14423, 2020. 2

[35] Diederik Paul Moeys, Federico Corradi, Emmett Kerr, Philip Vance, Gautham Das, Daniel Neil, Dermot Kerr, and Tobi Delbrück. Steering a predator robot using a mixed frame/event-driven convolutional neural network. In *EBCCSP*, pages 1–8. IEEE, 2016. 2

[36] Mohammad Mostafavi, Jonghyun Choi, and Kuk-Jin Yoon. Learning to super resolve intensity images from events. In *CVPR*, 2020. 2

[37] Mohammad Mostafavi, Lin Wang, and Kuk-Jin Yoon. Learning to reconstruct hdr images from events, with applications to depth and flow prediction. *IJCV*, pages 1–21. 1, 2

[38] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017. 2

[39] Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. Phased lstm: Accelerating recurrent network training for long or event-based sequences. In *NeurIPS*, 2016. 2

[40] 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

[41] Shivam Pande, Avinandan Banerjee, Saurabh Kumar, Biplab Banerjee, and Subhasis Chaudhuri. An adversarial approach to discriminative modality distillation for remote sensing image classification. In *CVPRW*, pages 0–0, 2019. 2

[42] Federico Paredes-Vallés and Guido CHE de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. *arXiv preprint arXiv:2009.08283*, 2020. 1

[43] SeongUk Park and Nojun Kwak. Feed: Feature-level ensemble for knowledge distillation. *arXiv preprint arXiv:1909.10754*, 2019. 2

[44] Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Emvs: Event-based multi-view stereo. 2016. 6

[45] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on Robot Learning*, pages 969–982, 2018. 2

[46] Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization. 2017. 3, 6

[47] Henri Rebecq, Timo Horstschäfer, Guillermo Gallego, and Davide Scaramuzza. Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real time. *RA-L*, 2(2):593–600, 2016. 6

[48] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *TPAMI*, 2019. 1, 2, 3, 4

[49] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2

[50] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *WACV*, pages 156–163, 2020. 2

[51] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *CVPR*, pages 1731–1740, 2018. 2

[52] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *ICCV*, pages 7244–7253, 2019. 2

[53] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. How to train your event camera neural network. *arXiv preprint*, 2020. 2

[54] 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 *ECCV*, 2020. 1, 2

[55] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *ICCV*, pages 1527–1537, 2019. 1, 2, 6

[56] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *arXiv preprint*, 2020. 2, 3

[57] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. *arXiv preprint*, 2020. 3

[58] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. *ECCV*, 2020. 1, 2, 3, 4, 8

[59] 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 *CVPR*, pages 608–619, 2021. 2

[60] 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 *CVPR*, pages 8315–8325, 2020. 2

[61] Lin Wang, S. Mohammad Mostafavi I., Yo-Sung Ho, and Kuk-Jin Yoon. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *CVPR*, pages 10081–10090, 2019. 1, 2, 3, 8

[62] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *TPAMI*, 2021. 2

[63] Lin Wang and Kuk-Jin Yoon. Semi-supervised student-teacher learning for single image super-resolution. *Pattern Recognition*, page 108206, 2021. 2

[64] Yanxiang Wang, Bowen Du, Yiran Shen, Kai Wu, Guangrong Zhao, Jianguo Sun, and Hongkai Wen. Ev-gait: Event-based robust gait recognition using dynamic vision sensors. In *CVPR*, pages 6358–6367, 2019. 2

[65] Yuanhao Wang, Ramzi Idoughi, and Wolfgang Heidrich. Stereo event-based particle tracking velocimetry for 3d fluid flow reconstruction. In *ECCV*, pages 36–53. Springer, 2020.

2

[66] Zihao W Wang, Peiqi Duan, Oliver Cossairt, Aggelos Katsaggelos, Tiejun Huang, and Boxin Shi. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In *CVPR*, pages 1609–1619, 2020. 1

[67] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, pages 675–684, 2018. 2, 3

[68] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *ECCV*, pages 588–604. Springer, 2020. 2

[69] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *CVPR*, pages 4968–4978, 2020. 2

[70] Anbang Yao and Dawei Sun. Knowledge transfer via dense cross-layer mutual-distillation. *ECCV*, 2020. 2

[71] Mingkuan Yuan and Yuxin Peng. Ckd: Cross-task knowledge distillation for text-to-image synthesis. *IEEE Transactions on Multimedia*, 2019. 2

[72] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 2

[73] Song Zhang, Yu Zhang, Zhe Jiang, Dongqing Zou, Jimmy Ren, and Bin Zhou. Learning to see in the dark with events. In *ECCV*, 2020. 2

[74] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018. 2

[75] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, pages 4106–4115, 2019. 3

[76] Long Zhao, Xi Peng, Yuxiao Chen, Mubbasir Kapadia, and Dimitris N Metaxas. Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. In *CVPR*, pages 6528–6537, 2020. 2, 5

[77] Yi Zhou, Guillermo Gallego, Henri Rebecq, Laurent Kneip, Hongdong Li, and Davide Scaramuzza. Semi-dense 3d reconstruction with a stereo event camera. In *ECCV*, pages 235–251, 2018. 6

[78] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *RA-L*, 3(3):2032–2039, 2018. 2, 5, 6

[79] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In *ECCV*, pages 711–714. Springer, 2018. 3, 8

[80] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*, pages 989–997, 2019. 1, 2, 3, 6, 7

[81] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 2020. 1, 2