EvDistill: Asynchronous Events to End-task Learning via Bidirectional Reconstruction-guided Cross-modal Knowledge Distillation

Lin Wang[1], Yujeong Chae[1]*, Sung-Hoon Yoon[1]*, Tae-Kyun Kim[2], and Kuk-Jin Yoon[1]
[1]Visual Intelligence Lab., KAIST, Korea
[2]ICVL Lab., KAIST, Korea and Imperial College London, UK
{wanglin, yujeong, yoon307, kimtaekyun, kjyoon}@kaist.ac.kr

Abstract

*Event cameras sense per-pixel intensity changes and produce asynchronous event streams with high dynamic range and less motion blur, showing advantages over the conventional cameras. A hurdle of training event-based models is the lack of large qualitative labeled data. Prior works learning end-tasks mostly rely on labeled or pseudo-labeled datasets obtained from the active pixel sensor (APS) frames; however, such datasets' quality is far from rivaling those based on the canonical images. In this paper, we propose a novel approach, called **EvDistill**, to learn a student network on the unlabeled and unpaired event data (target modality) via knowledge distillation (KD) from a teacher network trained with large-scale, labeled image data (source modality). To enable KD across the unpaired modalities, we first propose a bidirectional modality reconstruction (BMR) module to bridge both modalities and simultaneously exploit them to distill knowledge via the crafted pairs, causing no extra computation in the inference. The BMR is improved by the end-tasks and KD losses in an end-to-end manner. Second, we leverage the structural similarities of both modalities and adapt the knowledge by matching their distributions. Moreover, as most prior feature KD methods are uni-modality and less applicable to our problem, we propose an affinity graph KD loss to boost the distillation. Our extensive experiments on semantic segmentation and object recognition demonstrate that EvDistill achieves significantly better results than the prior works and KD with only events and APS frames.*

1. Introduction

Event cameras have recently received much attention in the computer vision and robotics community for their distinctive advantages, such as high dynamic range (HDR) and much less motion blur. Event cameras sense the intensity changes at each pixel asynchronously and produce event streams encoding time, pixel location, and polarity (sign) of intensity changes. Recently, deep neural network (DNN)-based methods with large-scale, labeled image data have shown significant performance gains on many tasks. However, learning effective event-based DNNs has been impeded by the lack of large pixel-level labeled event data. Prior works learning event-based high-level tasks have resorted to the manually annotated task-specific datasets in a supervised manner [1, 4, 7, 18, 19, 37, 40, 54, 62, 69]. Although some labeled event datasets [30, 45, 83] have been collected, the quantity and quality are far less favorable compared to those based on the canonical images. Some works [1, 18] have made pseudo labels using the active pixel sensor (APS) images; however, these labels are less accurate due to the low quality of APS images and considerable domain gap with the source data. While [83, 85] have explored unsupervised learning, they only focus on the pixel-level prediction tasks, *e.g.*, depth estimation. Another line of research has reconstructed intensity images from events [41, 54, 59, 63, 65, 66], and these images have been used to learn DNNs on end-tasks, *e.g.*, object recognition [54]; however, annotated labels are still needed, and extra latency is introduced in the inference time.

We explore to leverage large labeled image data (a.k.a. source modality) and the learned models, and aim to learn a model on the unpaired and unlabeled events (a.k.a. target

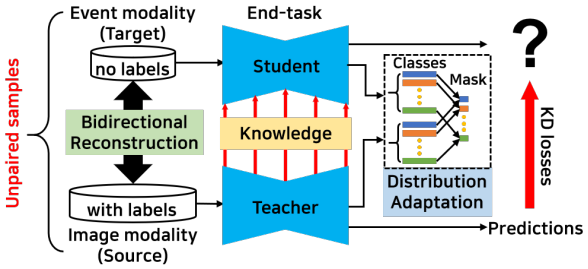

Figure 1: EvDistill distills knowledge from a teacher network trained with large labeled images to a student network learning unpaired and unlabeled events for the end-tasks. To distill knowledge, a bidirectional modality reconstruction and distribution adaptation schemes, with the novel KD losses, are proposed.

*These two authors contributed equally.

modality) via cross-modal learning [22, 81] and knowledge distillation (KD) [22, 61, 68]. Most existing cross-modal learning methods have relied on paired data (*e.g.*, image and depth) with the same labels [22, 23, 35, 46, 67, 71, 76, 81] or extra information (*e.g.*, data or labels) [2, 17, 50, 77] or grafting networks between modalities (*e.g.*, image to thermal) [29] for learning the end-tasks. Some works have explored the unpaired multi-modality data [13, 35]; however, it is assumed that labels for both modalities are available, which is difficult to achieve for the event data.

To overcome these limitations, we propose a novel method, called **EvDistill**, to efficiently learn a student network on the unpaired and unlabeled event data by distilling the knowledge from a robust teacher network trained with large labeled image data, as shown in Fig. 1. Firstly, we propose a novel bidirectional modality reconstruction (BMR) module to bridge both modalities, and then simultaneously exploit them to distill knowledge via the crafted pairs, adding no extra computation cost during inference (Sec. 3.2). Importantly, the BMR is improved by the end-task and the KD losses in an end-to-end manner. That is, BMR produces the crafted pairs of both modalities to distill knowledge to the student network in the *forward* pass, and KD facilitates the learning of the BMR in the *backward* pass. Secondly, as the feature representations of two modalities extracted from the task networks could suffer from distribution mismatch, we leverage the structural similarities and adapt knowledge by matching the class distributions based on the BMR module (Sec. 3.3). Moreover, as most existing feature KD methods are limited to uni-modality [33, 55, 78], we propose a novel graph affinity KD loss and other losses to learn a better model on the event data (Sec. 3.4). We evaluate the performance of the proposed framework on three datasets in semantic segmentation (Sec. 4.1) and one dataset in object recognition (Sec. 4.2). The experiments show that our approach achieves significantly better performance than the prior works for both end-tasks and the naive setting, KD with only events and the APS frames (when APS frames are available). The validation code and trained models are available at `https://github.com/addisonwang2013/evdistill`.

2. Related Works

DNNs for event-based end-tasks. DNNs with event data was first explored for the classification task [44] and for robot control [40]. [37] then trained a DNN for steering angle prediction on DDD17 dataset [5]. This dataset has been utilized by [1, 18] to perform semantic segmentation using pseudo labels obtained from the APS frames. Moreover, DNNs have been applied to some high-level prediction tasks, such as object detection and tracking [7, 29, 38, 51], human pose estimation [6, 69, 73], motion estimation [32, 39, 58, 70, 74], object recognition [4, 18, 54] on N-Caltech [45] and other datasets [4, 34, 57].

DNNs for event-based low-level vision. Meanwhile, another line of research focuses on the low-level prediction tasks, such as optical flow estimation [16, 19, 59, 85], stereo depth estimation [62, 85] on MVSEC dataset [83]. In addition, [42, 48, 54, 56, 59, 66] attempted to reconstruct intensity image/video from events using camera simulator [43, 52], and [41, 63, 65] tried to reconstruct high-resolution images. In contrast to image/video reconstruction from events, [18] proposed to generate events from video frames. Some other works also explored the potential of events for image deblurring [25, 31, 65], HDR imaging [24, 79], and event denoising [3]. For more details about event-based vision, refer to a survey [15]. Differently, we propose EvDistill, learning event-based end-tasks on the unpaired and unlabeled events via cross modal KD, in which a BMR moudle is proposed to bridge both modalities and is learned with the end-task networks in an end-to-end manner.

Knowledge Distillation. KD aims to build a smaller (student) model with the softmax labels of a larger (teacher) model [27, 55, 68]. Most KD methods learning end-tasks have been focused on uni-modality (*e.g.*, image) data and distill knowledge using logits [9, 72, 75, 80] or features [26, 33, 49, 55, 78]. Cross-modal KD aims to transfer knowledge across different modalities. Most prior cross-modal KD methods [8, 13, 22, 23, 28, 29, 35, 46, 47, 67, 71, 76, 81] relied on the paired data with the common labels, while some works utilized extra information (*e.g.*, data) [2, 17, 50, 77] or grafting networks [29] to transfer knowledge. Although [13, 35] explored the unpaired multi-modal KD, the labels for both modalities are needed. For more details about KD, refer to [68]. Learning from event data is more challenging as no paired labels for two modalities exist; we thus propose EvDistill, where we exploit a BMR module to connect them and simultaneously distill knowledge by adapting distribution with novel KD losses.

3. The Proposed Method: EvDistill

Event Representation 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. A natural choice is to encode events in a spatial-temporal 3D volume to a voxel grid [54, 84, 85] or event frame [19, 53] or multi-channel images [36, 63, 66]. In this paper, we represent events to multi-channel event images as the inputs to the DNNs. Details are provided in the suppl. material.

3.1. Overview

We describe the proposed EvDistill framework for learning end-tasks from events, as shown in Fig. 2. For event cameras, *e.g.*, DAVIS346 [60] with APS frames, assume that we are given the target modality data $\mathcal{X}_T = \{e, x_{aps}\}_i$

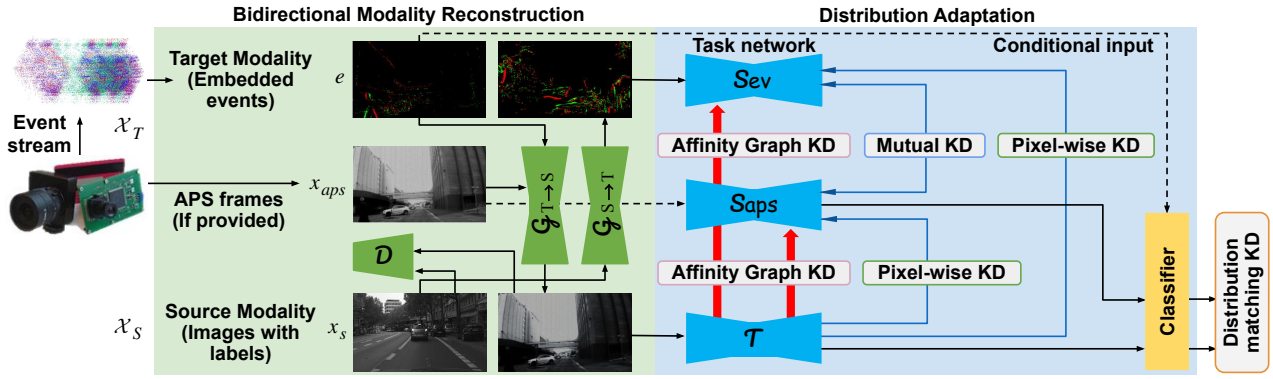

Figure 2: Overview of the proposed EvDistill framework. The architecture comprises a teacher network $\mathcal{T}$ and two student networks $\mathcal{S}_{ev}$ and $\mathcal{S}_{aps}$ (when APS frames exist). As there is no paired data with the same labels for both modality, a novel bidirectional reconstruction module is proposed to connect image and event modalities. Meanwhile, a distribution adaptation scheme with novel KD losses is also proposed to match the spatial structural distribution of both modalities.

without labels, where e_i and x_{aps_i} are i-th embedded event image and corresponding APS image. However, the source modality image data $\mathcal{X}_S = \{x_s, y_s\}_j$ are *with labels*, where x_{s_j} is j-th image with its label y_{s_j}. Unlike prior cross-modal learning methods [22, 81], we assume there are no paired source and target modality data with common labels. We address the challenge by proposing EvDistill, where $\mathcal{X}_S$ and $\mathcal{X}_T$ are not paired, and only the labels of $\mathcal{X}_S$ are available. In EvDistill, there are two student networks $\mathcal{S}_{ev}$ for learning events and $\mathcal{S}_{aps}$ for learning APS images (when APS images are provided). Our goal is to train a student network $\mathcal{S}_{ev}$ learning events by distilling knowledge from a teacher network $\mathcal{T}$. Our key ideas are three folds. First, as data of both modalities $\mathcal{X}_S$ and $\mathcal{X}_T$ are unpaired, we thus propose a bidirectional reconstruction module to bridge both modalities and then simultaneously exploit them to distill knowledge via the crafted pairs (Sec. 3.2). Second, as there exist spatial structure similarities (*e.g.*, cars, people in urban scenes) between the two modalities, we propose a distribution adaptation scheme to adapt the knowledge by matching the class distribution of the two modalities (Sec. 3.3). Lastly, as some end-tasks, *e.g.*, semantic segmentation, aim to predict pixel-wise class information, we propose a novel affinity graph KD loss and employ other loss terms to learn a better $\mathcal{S}_{ev}$ (Sec. 3.4).

3.2. Bridging Source and Target Modalities

Although data from both modalities are unpaired, we observe that one modality could be the alternative representation of the other modality under the same end-task. We thus propose a bidirectional modality reconstruction (BMR) module to bridge both modalities for enabling distillation on the $\mathcal{S}_{ev}$. As shown in Fig. 2, the proposed BMR consists of two generators where generator $\mathcal{G}_{T\to S}$ translates events to an intermediate representation in the image modality, and $\mathcal{G}_{S\to T}$ translates the labeled image data to an intermediate representation in the event modality. In such a way, the labels of image modality can be leveraged as supervision on the intermediate representation in the event modality when training $\mathcal{S}_{ev}$. Meanwhile, the intermediate image representation of events also helps to learn the knowledge (*e.g.*, predicted labels) from the teacher $\mathcal{T}$. When the APS frames are available in some event cameras, we utilize APS frames and apply the pixel-wise loss for the supervision of $\mathcal{G}_{T\to S}(e)$. The generated images are further adapted to the source data $\mathcal{X}_S$. The pixel-wise loss is defined as:

$$\mathcal{L}_{BMR}^{pw} = \mathbb{E}_{e,x_{aps}\sim\mathcal{X}_T}[||x_{aps} - \mathcal{G}_{T\to S}(e)||_1]. \quad (1)$$

Moreover, BMR is enhanced by the cycle consistency loss [86] and adversarial loss, which are crucial for the mapping of the two modalities. The adversarial loss (*e.g.*, from target to source modality) is formulated based on [20, 64]:

$$\mathcal{L}_{BMR}^{Adv} = \mathbb{E}_{e\sim\mathcal{X}_T}[1 - \log(\mathcal{D}(\mathcal{G}_{T\to S}(e))], \quad (2)$$

where $\mathcal{G}$ is the generator and $\mathcal{D}$ is the discriminator.

End-to-end learning. To better preserve semantic information, we exploit a novel dynamic semantic consistency (DSC) loss and build our framework based on adversarial learning [20, 66, 82], as shown in Fig. 3. The proposed DSC loss for BMR has three advantages: (1) the generated intermediate representation of events in image modality $\mathcal{G}_{T\to S}(e)$ becomes the optimal input of $\mathcal{T}$ and the generated intermediate representation of source images in event modality $\mathcal{G}_{S\to T}(x_s)$ becomes the optimal input of $\mathcal{S}_{ev}$; (2) the generated intermediate representations $\mathcal{G}_{T\to S}(e)$ and $\mathcal{G}_{S\to T}(x_s)$ both provide the supervision for $\mathcal{S}_{ev}$ based on the knowledge of $\mathcal{T}$ and labels of image data; (3) importantly, the BMR module is improved by these constraints on the end-tasks (*e.g.*, cross-entropy loss) in an end-to-end manner (see Fig. 3). That is, the KD loss promotes learning of the BMR module in the *backward* pass, and the BMR module produces crafted pairs to distill knowledge to the student network $\mathcal{S}_{ev}$ in the *forward* pass, which will be described in Sec. 3.3. *The BMR module can be removed after training, leading to no additional computation cost during*

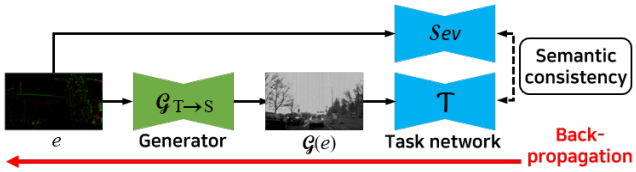

Figure 3: Illustration of the end-to-end target-to-source modality reconstruction $\mathcal{G}_{T\to S}$ with task net $\mathcal{T}$ and $\mathcal{S}_{ev}$ in the BMR module.

inference time. The proposed DSC loss for, *e.g.*, target-to-source modality learning is as:

$$\mathcal{L}_{BMR}^{DSC} = \mathbb{E}_{e\sim\mathcal{X}_T} KL[\mathcal{T}(\mathcal{G}_{T\to S}(e))||\mathcal{S}_{ev}(e)], \quad (3)$$

where $KL(\cdot||\cdot)$ is the KL divergence between two distributions. More detailed formulation of the proposed BMR and its total loss $\mathcal{L}_{BMR}$ is provided in the suppl. material.

3.3. Distillation via Distribution Adaptation (DA)

Based on the BMR module, the source and target modalities are connected, and we then further simultaneously exploit them to distill knowledge. Due to the distinct difference between event and image modality data, the features of two modalities extracted from the teacher and student networks suffer from the distribution mismatch. To address this issue, we propose to leverage the intrinsic spatial structure between source and target modality data $\mathcal{X}_S$ and $\mathcal{X}_T$. Our motivations are two folds. Firstly, as shown in Fig. 4, when APS frames are available, we propose to employ KD losses to guide the student $\mathcal{S}_{aps}$ to behave like the teacher $\mathcal{T}$ in addition to the cross-entropy (CE) loss $\mathcal{L}_{CE}$ based on source labels. This is done by aligning both $\mathcal{G}_{T\to S}(e)$ and x_{aps} with x_s to generalize the feature information. That is, the intermediate representation of events $\mathcal{G}_{T\to S}(e)$ and APS image x_{aps}, the source image x_s are all fed to the teacher $\mathcal{T}$ and the student $\mathcal{S}_{aps}$ to match the features. For simplicity, we model prediction matching loss based on the pixel-wise loss (*e.g.*, l_1), which is formulated as:

$$\begin{aligned}\mathcal{L}_{DA}^{aps} = \mathbb{E}_{x_{aps}\sim\mathcal{X}_T} KL[\mathcal{S}_{aps}(x_{aps})||\mathcal{T}(x_{aps})] + \\ \mathbb{E}_{x_s\sim\mathcal{X}_S} KL[\mathcal{S}_{aps}(x_s)||\mathcal{T}(x_s)]\end{aligned} \quad (4)$$

We then propose to employ a distillation loss to guide the student $\mathcal{S}_{ev}$ to behave more like the teacher $\mathcal{T}$ based on the event e, which can be formulated as:

$$\mathcal{L}_{DA}^{ev} = \mathbb{E}_{e,x_{aps}\sim\mathcal{X}_T}[KL[\mathcal{T}(x_{aps})||\mathcal{S}_{ev}(e))] \quad (5)$$

However, the source image-guided distillation can not fully reduce the distribution mismatch as pixels may vary in either the appearance or scales from $\mathcal{X}_T$ and $\mathcal{X}_S$. For instance, cars are always small, and buildings are always large in either event or image modality. We then propose to directly match the distribution of class categories between the two modalities, as shown in Fig. 4. Specifically, denote

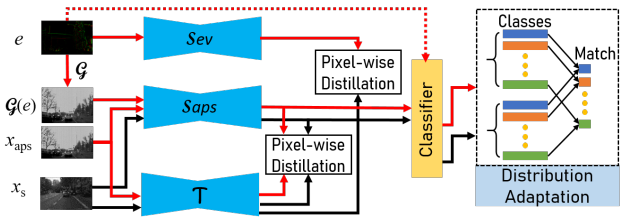

Figure 4: Illustration of distribution matching of the proposed distribution adaption scheme.

$h : x \to \{0, 1\}$ as a classifier, which is used to predict which modality an input pixel-level feature comes from, where 0 denotes the source modality $\mathcal{X}_S$, and 1 denotes the target modality $\mathcal{X}_T$. Intuitively, training a distribution classifier is to distinguish samples from two modalities. We encourage the activation x to be modality indistinguishable. Considering each x is generated from a task network, denoted by $\mathcal{F}$ (either $\mathcal{T}$ or $\mathcal{S}_{ev}$/$\mathcal{S}_{aps}$), we thus need to optimize $\mathcal{F}$ such that the distribution classification loss $\mathcal{L}_{DA}^{match}$ is maximized. By jointly learning the distribution classifier h and the task network $\mathcal{F}$, we arrive at the following maxmin problem, which can be optimized in an adversarial training manner [20].

$$\mathcal{L}_{DA}^{match}(\mathcal{X}_S, \mathcal{X}_T) = \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \mathcal{H}(h(x), d) \quad (6)$$

Here, $\mathcal{X} = \mathcal{X}_S \cup \mathcal{X}_T$, $|\mathcal{X}|$ is the number of samples, $d \in \{0, 1\}$ is the modality label, and $\mathcal{H}(\cdot)$ is a classification loss. For convenience, we adopt the conditional adversarial learning [12, 21] to optimize the matching problem. Finally, the distribution adaptation (DA) loss $\mathcal{L}_{DA}$ is the linear combination of the three loss terms $\mathcal{L}_{DA}^{aps}$, $\mathcal{L}_{DA}^{ev}$ and $\mathcal{L}_{DA}^{match}$.

3.4. Affinity Graph KD and Other KD Losses

Affinity Graph (AG) KD. Teacher's features contain constructive knowledge; however, due to modality difference, directly matching feature information [49, 81] is impractical. We notice that two modalities share similar labeling contiguity among spatial locations for, *e.g.*, urban scenes. We thus build affinity graphs to transfer the instance-level similarity along the spatial locations between two modalities, as shown in Fig. 5. The node represents a spatial location of an instance (*e.g.*, car), and the edges connected between two nodes represent 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. 5) 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. In the two modalities, we denote $A_{uv}^{\mathcal{T}}$ and $A_{uv}^{\mathcal{S}}$ are the affinities between the u-th node

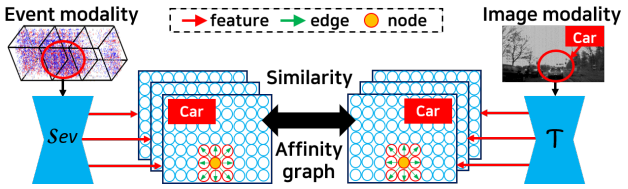

Figure 5: Illustration of the proposed affinity graph distillation loss between the event and image modalities.

and the v-th node obtained from the teacher and student, respectively, which is formulated as:

$$\mathcal{L}_{AG} = \frac{1}{H \times W \times \sigma} \sum_{u \sim R} \sum_{v \sim \sigma} ||A^{\mathcal{T}}_{uv} - A^{\mathcal{S}}_{uv}||_2^2 \quad (7)$$

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_u and F_v as $A_{uv} = \frac{F_u^\intercal F_v}{||F_u\intercal||_2||F_v||_2}$, where $F_u^\intercal$ is the transposed feature vector of F_u.

Mutual Distillation (MD). When APS frames exist, $\mathcal{S}_{aps}$ indeed can facilitate the learning of $\mathcal{S}_{ev}$. Since $\mathcal{S}_{aps}$ and $\mathcal{S}_{ev}$ start from different initial conditions, they learn different representations, and consequently, their prediction of probabilities can be an effective regularization to each other. We thus let $\mathcal{S}_{aps}$ and $\mathcal{S}_{ev}$ learn from each other's predictions via the KL divergence losses with a temperature parameter τ [80]. This helps $\mathcal{S}_{ev}$ converge to better minima for better generalization to test data. The MD loss is formulated as:

$$\begin{aligned}\mathcal{L}_{MD} = \mathbb{E}_{e,x_{aps} \sim \mathcal{X}_T} KL\left[\mathcal{S}_{ev}(e) | \mathcal{S}_{aps}(x_{aps}), \tau\right] + \\ \mathbb{E}_{e,x_{aps} \sim \mathcal{X}_T} KL\left[\mathcal{S}_{aps}(x_{aps}) | \mathcal{S}_{ev}(e), \tau\right].\end{aligned} \quad (8)$$

In summary, the objective $\mathcal{L}$ of EvDistill is as follows :

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{BMR} + \lambda_1 \mathcal{L}_{DA} + \lambda_2 \mathcal{L}_{AG} + \lambda_3 \mathcal{L}_{MD} \quad (9)$$

where λ_1, λ_2 and λ_3 are the hyper-parameters.

4. Experiment and Evaluation

4.1. Event-based semantic segmentation

Semantic segmentation is a task that aims to assign a semantic label, *e.g.*, road, car, in a given scene to each pixel. Unlike image data, annotating events requires special processing for the raw event streams. Besides, it is challenging to correctly label the pixels due to the sparsity of events and lacking information (*e.g.*, material and textures).

Datasets. We use the publicly available driving scene dataset DDD17 [5], 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). However, such a way leads to less precise segmentation labels because there is a considerable domain gap between the source data and APS images of low resolution and quality. Note that our method does not rely on any annotations of events in training, and the pseudo annotations by [1] for test are only used for evaluation and comparison. As the events in the DDD17 dataset are very sparse and noisy, we show more results on the driving sequences in the MVSEC dataset [85], collected for the stereo purpose. Moreover, we show qualitative results on the E2Vid driving scene dataset [54] captured by using a Samsung event camera (with higher resolution).

Implementation details. For each label of DDD17 dataset provided by [1], we use the events occurred in a 50ms time window before a label for prediction, as done in [1, 18]. We also consider the event representation method in [1] for semantic segmentation on the DDD17 dataset, in addition to that described in Sec. 3. For the event representation on MVSEC and E2VID datasets, we use the method in Sec. 3. For the teacher $\mathcal{T}$ and students $\mathcal{S}_{ev}$ and $\mathcal{S}_{aps}$, we adopt a segmentation network [11]. We use the following metric to evaluate the performance. The *intersection of union* (IoU) score is calculated as the ratio of intersection and union between the GT mask and the predicted segmentation mask for each class. We use the *mean IoU* (MIoU) to measure the effectiveness, as done in [10, 11]. The segmentation maps are with 6 classes, as done in [1]. For more implementation details (*e.g.*, training), refer to the suppl. material.

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 existing methods, EvSegNet [1] and Vid2E [18] that uses synthetic version of DDD17 data. It turns out that, without using labels, EvDistill significantly improves the segmentation results on events and surpasses the existing methods with around 7% increase in MIoU using a multi-channel event representation. Segmentation using the voxel grid representation is slightly less effective than that of using the multi-channel representation. Meanwhile, on the time interval of 10ms and 250ms, EvDistill also shows a significant increase of MIoU by around 7.5% and 9.5% than those of EvSegNet, respectively.

We also demonstrate that our method significantly enhances the semantic segmentation performance on the APS frames in Table 1. Quantitatively, without resorting to the pseudo labels, our method surpasses EvSegNet by a large margin with around 12% increase of MIoU. This reflects that our method dramatically minimizes the domain gap between APS frames and labeled source data and distills the knowledge from the teacher network to learn a student network $\mathcal{S}_{aps}$ showing better segmentation capability. The results on both events and APS frames indicate that our method successfully distills the knowledge from the teacher

Table 1: Segmentation performance with different event representations and APS images on the test data [1], measured by MIoU.

Method	Event Rep.	Use pseudo labels	MIoU [50ms]	MIoU [10ms]	MIoU [250ms]
EvSegNet [1]	6-channel [1]	Yes	54.81	45.85	47.56
Vid2E [18]	EST [19]	Yes	45.48	30.70	40.66
Ours	Voxel bins (2ch)	No	57.16	48.68	51.23
Ours	Multi-channel	No	**58.02**	**49.21**	**52.01**
EvSegNet (APS)	-	Yes	64.98	64.98	64.98
Ours (APS)	-	No	**72.63**	**72.63**	**72.63**

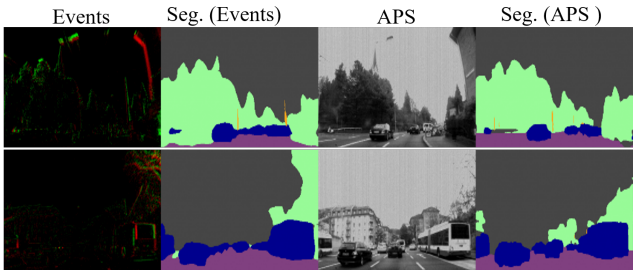

Figure 6: Semantic segmentation results of urban driving scenes on DDD17 test dataset (gray: background; green: vegetation; blue: vehicle; violet: street; yellow: object).

network learned on the labeled image modality data for tackling the challenges of unpaired and unlabeled events.

High dynamic range (HDR). HDR is one distinct advantage of event cameras. Even when APS frames are ill-exposed, events capture the intensity changes. We show the student network $\mathcal{S}_{ev}$ shows promising performance in the extreme condition. Fig. 4 of the suppl. material shows the qualitative results. The APS frames are over-exposed, thus the student network $\mathcal{S}_{aps}$ fails to segment the urban scenes; however, the events capture the scene details, and the student network $\mathcal{S}_{ev}$ shows convincing segmentation results.

4.1.2 Evaluation on MVSEC dataset

We further present the experimental results on the MVSEC dataset [85], which contains various driving scenes for stereo estimation. We use the 'outdoor_day2' sequence and divide the data into training and test sets. 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 for HDR. For the details of dataset preparation and more visual results, refer to the suppl. material. To quantitatively evaluate our method, we also utilize the APS frames to generate pseudo labels, similar to [1], as our comparison baseline. The qualitative and quantitative results are shown in Fig. 7 (also see Fig. 2 of the suppl. material) and Table 2. In Fig. 7, we mainly show the results in the low-light condition. It is evident that, although the student network $\mathcal{S}_{aps}$ fails to work on APS frames, where most pixels in the red boxes are wrongly classified in the 4th column (*e.g.*, building and trees misclassified as vehicles), events capture scene information better and provide better segmentation performance in low-light condition. Compared with the baseline in Table 2, our method significantly surpasses it by a noticeable margin with a 8.8% increase of MIoU on the events and a 11.2% increase on the APS frames.

Table 2: Segmentation performance of events and APS images on the MVSEC dataset [83], measured by MIoU. The baseline is trained by using the pseudo labels made from the APS images.

Method	Use pseudo labels	MIoU
Baseline (Events)	Yes	50.53
Ours(Events)	No	**55.09**
Baseline (APS)	Yes	61.93
Ours (APS)	No	**68.85**

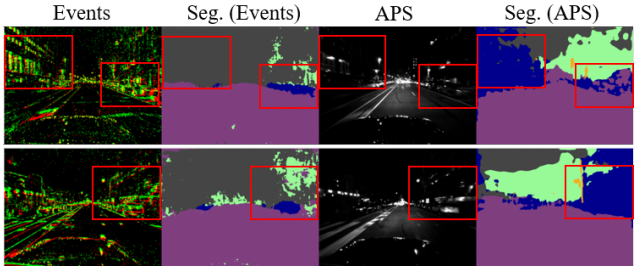

Figure 7: Semantic segmentation results of low-light scenes on MVSEC dataset (gray: background; green: vegetation; blue: vehicle; violet: street; yellow: object).

4.1.3 Evaluation on E2VID dataset

We also present the experimental results on the E2Vid driving dataset [54]. We followed the DDD17 in [1] to split E2VID dataset. We mainly use 'sun2' and 'sun4' sequences where we select around 4K event images as the training set, and the remained 400 as the test set. We also test on 400 event images from the 'street' sequence. As there are no GT annotations for events, we only show the qualitative results, as shown in Fig. 8 and Fig. 3 of the *suppl. material*. The student network $\mathcal{S}_{ev}$ can segment the moving objects, *e.g.*, vehicles, pedestrians. Meanwhile, it also successfully segments the complex objects with no motion blur,*e.g.*, tree branches, traffic lights. Compared with the events (346x260) in DDD17 and MVSEC datasets, the events in E2Vid are in a higher resolution (640x480). In Fig. 8, it seems that the network trained on these events better segments small objects, *e.g.*, vehicles in the remote location, traffic lights, etc. Although events contain less visual information than the image data, it is advantageous for segmenting the fast moving objects and HDR scenes.

4.2. Event-based object recognition

We further demonstrate that our method can be also flexibly applied to object recognition. We use the benchmark N-Caltech101 dataset [45]. This dataset is an event-based

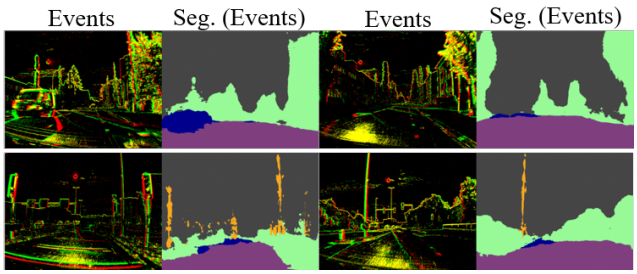

Figure 8: Qualitative results on E2Vid dataset (gray: background; green: vegetation; blue: vehicle; violet: street; yellow: object).

Table 3: A comparison of object recognition performance with existing methods on N-Caltech dataset.

Method	Training data	Use GT labels	Test score
HATS [57]	Real events	Yes	0.642
HATS-ResNet34	Real events	Yes	0.691
RG-CNN	Real events	Yes	0.657
EST [19]	Real events	Yes	0.817
E2VID [54]	Intensity images	Yes	0.866
VID2E [18]	Synthetic events	Yes	0.807
Ours-20K events	Real events	No	0.896
Ours (fine-tune)	Real events	No	**0.902**

version of the well-known Caltech101 [14]. Note that the event data in the N-Caltech dataset and the original images in the Caltech dataset are not matched. As the dataset only provides events captured by an event camera, without APS frames, the student network $\mathcal{S}_{aps}$ depicted in Fig. 2 is removed in this case. To bridge both modalities, we explore the source images (*e.g.*, the images from Caltech dataset and other images) and learn the bidirectional modality reconstruction (BMR) module in an unsupervised manner. Note that we assume that the labels of event data are unknown and the labels of source images are given. We utilize a teacher classifier $\mathcal{T}$ pretrained on the source images and distill the knowledge to the student classifier $\mathcal{S}_{ev}$. Interestingly, EvDistill translates the source images (with labels) to events also with the same labels (Fig. 10). We then utilize the generated events and target events to train the student classifier $\mathcal{S}_{ev}$. Meanwhile, the generated images (Fig. 9) and source images are used to get recognition information from the teacher $\mathcal{T}$, such that the student can learn the distilled knowledge from the teacher via the KD losses. From the experiments, we show there is a significant performance boost (see Table 3) with our method.

Implementation Details. For the teacher and student classifiers, we use the ResNet34, as used in other works [18, 19, 54]. We use the loss functions defined in Eq. (3) and Eq. (7), and also cross-entropy loss. We use Adam optimizer with the learning rate of 1e-4. For event representation, we use the stacking method described in Sec. 3. Due to the lack of space, more details of the implementation are provided in the supplementary material.

Experimental results. The qualitative results for image

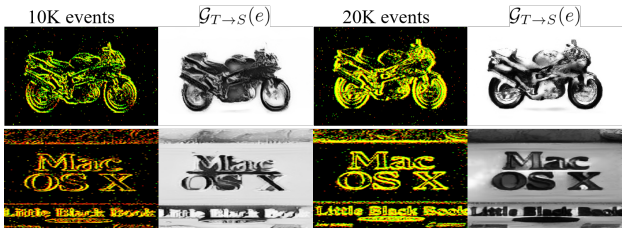

Figure 9: Generated images with $\mathcal{G}_{T\to S}$ (2nd and 4th columns based on 10K and 20K events) from the target events.

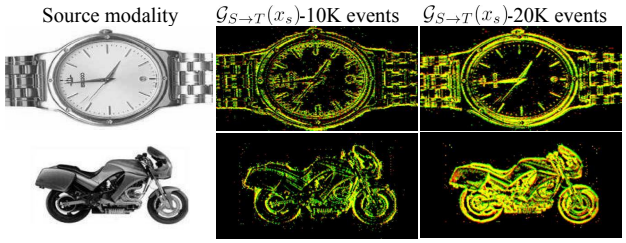

Figure 10: Generated events with $\mathcal{G}_{S\to T}$ (10K and 20K events in the 2nd and 3rd columns) from the source images.

generation are shown in Fig. 9. The first and third columns show the stacked 10K and 20K events, accompanied by the generated images in the 2nd and 4th columns. As can be visually seen, even without supervision, realistic images are generated from the target modality. When more events are accumulated, better-reconstructed images are obtained. Correspondingly, Fig. 10 shows the generated events from the source modality. The 1st column shows the source images, and the 2nd and 3rd columns are the generated 10K and 20K events, respectively. It is clearly shown that the quality of the generated events is improved based on the target events. Regarding the results of object recognition, we compare with the SoTA optimization-based methods, HATS [57] and its ResNet34-based result. Meanwhile, we compare with the recent DL-based methods, EST [19], E2Vid [54] (using generated images) and Vid2E [18] (using synthetic events). Table 3 shows the quantitative results. Even when labels are not used for event data, our method surpasses the existing methods with a significant margin. For instance, compared with HATS-Resnet34, our method achieves more than 20% higher accuracy. When compared with E2Vid, our method also achieves better results (0.896 vs 0.866). When tuning parameters via self-ensemble in training, EvDistill further improves the performance and achieves higher accuracy (0.902 vs 0.896).

5. Ablation Study and Analysis

Modality reconstruction. We show that EvDistill also enhances the target-to-source reconstruction by leveraging the proposed dynamic semantic consistency (DSC) KD loss. The qualitative and quantitative results are shown in Fig. 11 and Table 4. In contrast to the reconstructed images without KD loss (2nd column), our method fully exploits the semantic information and successfully restores the textural

Table 4: Comparison of segmentation results on the target to source reconstruction with and without knowledge distillation.

Method	Mean IoU
w/o knowledge distillation	43.64
w/ knowledge distillation	**45.12**

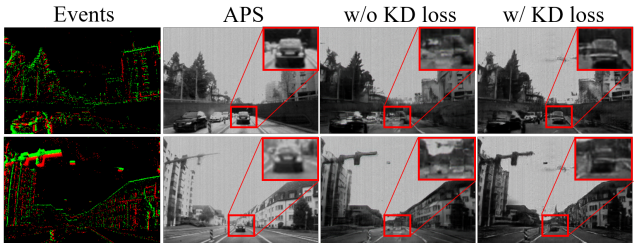

Figure 11: Visual results of the end-to-end target-to-source reconstruction with (w/) and without (w/o) KD loss.

and material details, *e.g.*, cars in the cropped patches (3rd column) in Fig. 11. The effectiveness can be further verified from Table 4, where the images generated with the KD loss show higher performance for semantic segmentation. The proposed EvDistill helps recover the semantic information in the generated images and improves semantic segmentation quality on these images.

The effectiveness of affinity graph KD. To further validate the effectiveness of the proposed affinity graph (AG) KD for cross-modal learning, we compare with the general feature-level KD losses, *e.g.*, FitNets [55], AT [78], and FT [33]. The results are shown in Table 5. As FitNets, AT and FT are all targeted to directly minimizing feature difference between the teacher and student under the same modality (*e.g.*, image) data, they show relatively poor performance on the cross-modal learning problem. Instead of directly matching features, the proposed AG loss better tackles the spatial contiguity of instances between the two modalities and shows better performance on the end-tasks.

The effectiveness of distillation. We look into the effect of enabling and disabling different components of the proposed EvDistill. The experiments were conducted on the semantic segmentation task with the DDD17 dataset. In Table 6, the results of different settings for the student network are listed. Our baseline framework consists of a teacher $\mathcal{T}$ and two student networks. From Table 6, we can see that KD can improve the performance of both student networks. Moreover, each KD scheme leads to higher test scores. This implies that the KD schemes make a complementary contribution to learning better student network. Furthermore, it is shown that the distribution matching KD scheme better matches the structural similarities between the modalities, leading to higher scores and better quality. On the other hand, the proposed BMR approach (with (w/) BMR) also contributes to enhancing the learning of the event-based segmentation network. When the student network $\mathcal{S}_{aps}$ is added, it also benefits the learning of $\mathcal{S}_{ev}$ optimized by the mutual learning KD and distribution matching KD losses.

Table 5: A comparison of affinity graph KD with existing feature KD methods on DDD17 dataset.

Metric	Event Rep.	FitNet [55]	AT [78]	FT [33]	AG
MIoU	Voxel-2ch	55.09	55.60	55.51	57.16

Table 6: The effect of different components of EvDistill with a multi-channel event representation. PI: pixel-wise KD, AG: affinity graph, DM: distribution adaptation, ML: mutual learning.

Method	Use pseudo labels	Performance (MIoU)
PI	No	55.10
PI + AG	No	56.59
PI + AG + DM	No	57.86
PI + AG + DM + ML	No	58.02
w/o BMR	No	56.25
w/ BMR	No	57.40
w/ BMR + $\mathcal{S}_{aps}$	No	58.02

KD with only events and APS frames. Although the paired events and APS frames are without annotated labels, one naive way might be to utilize them in EvDistill without exploring source data and the BMR module. We study this baseline by feeding the events to the student $\mathcal{S}_{ev}$ and the APS frames to the teacher $\mathcal{T}$ for semantic segmentation on the MVSEC dataset. We apply the proposed distribution adaptation scheme and the affinity graph loss to distill knowledge to the student $\mathcal{S}_{ev}$. The experimental results show that it achieves less plausible MIoU (52.10 vs. 55.09) than the proposed framework as there is a domain gap with the source data used to train the teacher network. The APS frames are of low quality, which degrades the performance.

6. Conclusion, Limitations and Future Work

This paper proposed EvDistill to learn a student on the unpaired and unlabeled events by distilling the knowledge from a teacher trained with labeled images. As no paired modality data with common labels exist, we proposed a BMR module to bridge both modalities. We also proposed a distribution adaptation scheme to match the distributions of two modalities. Besides, a novel graph affinity KD was proposed to enhance the KD performance. The experiments on two end-tasks demonstrate the effectiveness of our method. Our work has some limitations. First, the type of event data (*e.g.*, urban driving) needs to be close to the labeled source data. Second, as deep network is used, inference latency is inevitable. As EVDistill is a general approach, which tackles the problem caused by limited labels in the target modality. Thus, it can be flexibly applied to any other data, such as thermal and depth camera data in the future work.

Acknowledgement This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2018R1A2B3008640) and Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2014-3-00123, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis).

References

[1] Inigo Alonso and Ana C Murillo. Ev-segnet: semantic segmentation for event-based cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2, 5, 6

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6131, 2020. 2

[3] R Baldwin, Mohammed Almatrafi, Vijayan Asari, and Keigo Hirakawa. Event probability mask (epm) and event denoising convolutional neural network (edncnn) for neuromorphic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1701–1710, 2020. 2

[4] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based object classification for neuromorphic vision sensing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 491–501, 2019. 1, 2

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

[6] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[7] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. Asynchronous convolutional networks for object detection in neuromorphic cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2

[8] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 865–872, 2019. 2

[9] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. *arXiv preprint arXiv:1912.00350*, 2019. 2

[10] 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. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 5

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 5

[12] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2018. 4

[13] Qi Dou, Quande Liu, Pheng Ann Heng, and Ben Glocker. Unpaired multi-modal segmentation via knowledge distillation. *arXiv preprint arXiv:2001.03111*, 2020. 2

[14] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 7

[15] 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. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2020. 2

[16] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12280–12289, 2019. 2

[17] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 2

[18] 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. 1, 2, 5, 6, 7

[19] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5633–5643, 2019. 1, 2, 6, 7

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3, 4

[21] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017. 4

[22] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016. 2, 3

[23] 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

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1730–1739, 2020. 2

[25] 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

[26] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1921–1930, 2019. 2

[27] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distill-

ing the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[28] Hengtong Hu, Lingxi Xie, Richang Hong, and Qi Tian. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3123–3132, 2020. 2

[29] 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

[30] Yuhuang Hu, Hongjie Liu, Michael Pfeiffer, and Tobi Delbruck. Dvs benchmark datasets for object tracking, action recognition, and object recognition. *Frontiers in neuroscience*, 10:405, 2016. 1

[31] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3329, 2020. 2

[32] Daniel R Kepple, Daewon Lee, Colin Prepsius, Volkan Isler, and Il Memming. Jointly learning visual motion and confidence from local patches in event cameras. *European Conf. Comput. Vis.(ECCV)*, 2020. 2

[33] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *Advances in Neural Information Processing Systems*, pages 2760–2769, 2018. 2, 8

[34] Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11:309, 2017. 2

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

[36] 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 *European Conference on Computer Vision*, volume 3, 2020. 2

[37] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5419–5427, 2018. 1, 2

[38] Nico Messikommer, Daniel Gehrig, Antonio Loquercio, and Davide Scaramuzza. Event-based asynchronous sparse convolutional networks. *ECCV*, 2020. 2

[39] Anton Mitrokhin, Zhiyuan Hua, Cornelia Fermuller, and Yiannis Aloimonos. Learning visual motion segmentation using event surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14414–14423, 2020. 2

[40] 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 *2016 Second International Conference on Event-based Control, Communication, and Signal Processing (EBCCSP)*, pages 1–8. IEEE, 2016. 1, 2

[41] Mohammad Mostafavi, Jonghyun Choi, and Kuk-Jin Yoon. Learning to super resolve intensity images from events. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, 2020. 1, 2

[42] 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*, pages 1–21. 2

[43] 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

[44] Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. Phased lstm: Accelerating recurrent network training for long or event-based sequences. In *Advances in neural information processing systems*, pages 3882–3890, 2016. 2

[45] 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. 1, 2, 6

[46] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer, 2016. 2

[47] Shivam Pande, Avinandan Banerjee, Saurabh Kumar, Biplab Banerjee, and Subhasis Chaudhuri. An adversarial approach to discriminative modality distillation for remote sensing image classification. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[48] 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. 2

[49] SeongUk Park and Nojun Kwak. Feed: Feature-level ensemble for knowledge distillation. *arXiv preprint arXiv:1909.10754*, 2019. 2, 4

[50] Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Domain transfer for 3d pose estimation from color images without manual annotations. In *Asian Conference on Computer Vision*, pages 69–84. Springer, 2018. 2

[51] Bharath Ramesh, Shihao Zhang, Zhi Wei Lee, Zhi Gao, Garrick Orchard, and Cheng Xiang. Long-term object tracking with a moving event camera. In *Bmvc*, page 241, 2018. 2

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

[53] Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization. 2017. 2

[54] 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*, 2019. 1, 2, 5, 6, 7

[55] 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, 8

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

[57] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1731–1740, 2018. 2, 7

[58] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7244–7253, 2019. 2

[59] 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 Conf. Comput. Vis.(ECCV)*, 2020. 1, 2

[60] Gemma Taverni, Diederik Paul Moeys, Chenghan Li, Celso Cavaco, Vasyl Motsnyi, David San Segundo Bello, and Tobi Delbruck. Front and back illuminated dynamic and active pixel vision sensors comparison. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(5):677–681, 2018. 2

[61] Fida Mohammad Thoker and Juergen Gall. Cross-modal knowledge distillation for action recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 6–10. IEEE, 2019. 2

[62] 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 International Conference on Computer Vision*, pages 1527–1537, 2019. 1, 2

[63] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. *European Conf. Comput. Vis.(ECCV)*, 2020. 1, 2

[64] Lin Wang, Wonjune Cho, and Kuk-Jin Yoon. Deceiving image-to-image translation networks for autonomous driving with adversarial perturbations. *IEEE Robotics and Automation Letters*, 5(2):1421–1428, 2020. 3

[65] 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. 1, 2

[66] Lin Wang, Mostafavi I. S. Mohammad, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10081–10090, 2019. 1, 2, 3

[67] Lichen Wang, Jiaxiang Wu, Shao-Lun Huang, Lizhong Zheng, Xiangxiang Xu, Lin Zhang, and Junzhou Huang. An efficient approach to informative feature extraction from multimodal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5281–5288, 2019. 2

[68] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[69] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6358–6367, 2019. 1, 2

[70] Yuanhao Wang, Ramzi Idoughi, and Wolfgang Heidrich. Stereo event-based particle tracking velocimetry for 3d fluid flow reconstruction. In *European Conference on Computer Vision*, pages 36–53. Springer, 2020. 2

[71] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018. 2

[72] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *European Conference on Computer Vision*, pages 588–604. Springer, 2020. 2

[73] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4968–4978, 2020. 2

[74] Sifan Yang, Qi Zheng, Xiaowei Hu, and Guijin Wang. Vess: Variable event stream structure for event-based instance segmentation benchmark. In *Proceedings of the 2020 4th International Conference on Digital Signal Processing*, pages 112–116, 2020. 2

[75] Anbang Yao and Dawei Sun. Knowledge transfer via dense cross-layer mutual-distillation. *European Conference on Computer Vision*, 2020. 2

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

[77] Shanxin Yuan, Bjorn Stenger, and Tae-Kyun Kim. Rgb-based 3d hand pose estimation via privileged learning with depth images. *arXiv preprint arXiv:1811.07376*, 2018. 2

[78] 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, 8

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

[80] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018. 2, 5

[81] Long Zhao, Xi Peng, Yuxiao Chen, Mubbasir Kapadia, and Dimitris N Metaxas. Knowledge as priors: Cross-modal knowledge generalization for datasets without supe-

rior knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6528–6537, 2020. 2, 3, 4

[82] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 7287–7300, 2019. 3

[83] 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. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. 1, 2, 6

[84] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In *European Conference on Computer Vision*, pages 711–714. Springer, 2018. 2

[85] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 1, 2, 5, 6

[86] 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. 3