EventDance: Unsupervised Source-free Cross-modal Adaptation for Event-based Object Recognition

Xu Zheng[1] Lin Wang[1,2*]
[1]AI Thrust, HKUST(GZ) [2]Dept. of CSE, HKUST
zhengxu128@gmail.com, linwang@ust.hk

Project Page: https://vlislab22.github.io/EventDance/

Abstract

*In this paper, we make the **first** attempt at achieving the cross-modal (*i.e.*, image-to-events) adaptation for event-based object recognition **without accessing** any labeled source image data owning to privacy and commercial issues. Tackling this novel problem is non-trivial due to the novelty of event cameras and the distinct modality gap between images and events. In particular, as only the source model is available, a hurdle is how to extract the knowledge from the source model by only using the unlabeled target event data while achieving knowledge transfer. To this end, we propose a novel framework, dubbed **EventDance** for this unsupervised source-free cross-modal adaptation problem. Importantly, inspired by event-to-video reconstruction methods, we propose a reconstruction-based modality bridging (**RMB**) module, which reconstructs intensity frames from events in a self-supervised manner. This makes it possible to build up the surrogate images to extract the knowledge (*i.e.*, labels) from the source model. We then propose a multi-representation knowledge adaptation (**MKA**) module that transfers the knowledge to target models learning events with multiple representation types for fully exploring the spatiotemporal information of events. The two modules connecting the source and target models are mutually updated so as to achieve the best performance. Experiments on three benchmark datasets with two adaption settings show that EventDance is on par with prior methods utilizing the source data.*

1. Introduction

Event cameras, *a.k.a.*, the silicon retina [47], are bio-inspired novel sensors that perceive per-pixel intensity changes asynchronously and produce event streams encoding the time, pixel position, and polarity (sign) of the intensity changes [20, 21, 28]. Event cameras possess distinct merits, *e.g.*, high dynamic range and no motion blur, making them more advantageous for challenging visual conditions, where the sensing quality degrades for the frame-based cameras. As a result, event cameras have recently drawn much attention from the computer vision and robotics community [4, 6, 38, 48, 55, 70, 78]. Although events are sparse and mostly encode the edge information, it has been shown that events alone are possible for learning scene understanding tasks, *e.g.*, object recognition [20], via deep neural networks (DNNs). However, learning event-based DNNs is often impeded by the lack of large-scale precisely annotated datasets due to the asynchronous and sparse properties of event cameras, making it hardly possible to apply now-straightforward supervised learning. For these reasons, some research endeavors have been made to explore the cross-modal adaptation to transfer knowledge from the labeled image modality (*i.e.*, source) to the unlabeled event modality (*i.e.*, target) [40, 57, 68].

In this paper, we make the **first** attempt to achieve cross-modal (*i.e.*, image-to-event) adaptation for event-based ob-

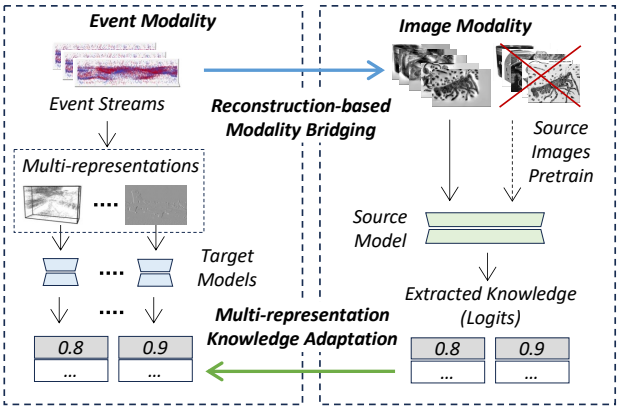


Figure 1. Illustration of the challenging task of the cross-modal adaptation from image to event modalities. We address it by introducing reconstruction-based modality bridging and multi-representation knowledge adaptation modules.

*Corresponding author.

ject recognition where ***we have no access to any labeled source image***. The assumption of such a novel problem is of great importance for conditions where the labeled images, *i.e.*, source modality data, *are not allowed to be released* due to privacy and commercial issues [1, 2, 33, 46], and *only the trained source models are shared*. However, tackling this problem is arduous due to 1) the sparse and asynchronous properties of events, making it difficult to directly apply existing cross-modal adaptation techniques,*e.g.*, [3], and 2) the significant modality gap between images and events as events mostly reflect edge information. In particular, as only the source modal is available, a hurdle is ***how to extract the knowledge* (i.e., *pseudo label) *from the source model by only using the unlabeled events and achieving knowledge transfer***.

To this end, we formulate the learning objectives as **1)** bridging the modality gaps between image and event modalities and **2)** transferring the knowledge from the source image model to the target event domain. Accordingly, we propose a novel unsupervised source-free cross-modal adaptation framework, dubbed **EventDance**. Importantly, inspired by the attempt for event-to-video reconstruction [44], we first propose a reconstruction-based modality bridging (**RMB**) module (Sec. 3.2), which builds up a **surrogate** image domain to imitate the source image distribution in a self-supervised manner. This allows for mitigating the large modality gap between images and events. Specifically, the RMB module takes an event stream to reconstruct multiple intensity frames to construct surrogate data in the image modality, which connects the source domain (inaccessible) of image modality. *However, using [44] alone does not meet our needs as it only aims to reconstruct natural-looking images, not optimal surrogate images (inputs) for the source model*. Therefore, we optimize our RMB module for better knowledge extraction (*i.e.*, pseudo labels) *by minimizing the entropy of the source model's prediction* and ensuring *temporal consistency* based on the high temporal resolution of event data.

Buttressed by the surrogate domain, we then propose a multi-representation knowledge adaptation (**MKA**) module (Sec.3.3) that transfers the knowledge to learn target models for unlabeled events. As significant information loss, *e.g.*, timestamp drops, occurs when converting events to a specific representation, like event count image, it may hinder the object recognition performance [70]. Therefore, we leverage several event representations in our EventDance, including stack image [32], voxel grid [63], and event spike tensor (EST) [21], to fully explore the spatio-temporal information of events. *This allows for maintaining the cross-model prediction consistency training between the target models*. These two modules connecting the source and target models are mutually updated so as to achieve better modality bridging and knowledge adaptation.

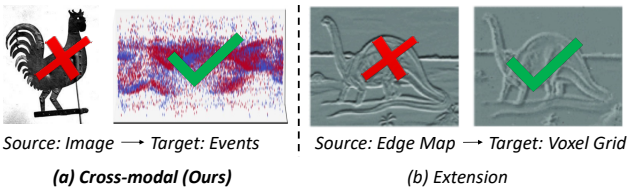


Figure 2. **Different adaptation settings**. (a) ours from image to event modalities. (b) SFUDA from different image types [68].

We validate EventDance on three event-based recognition benchmarks: N-Caltech 101 [43], N-MNIST [43], and CIFAR10-DVS [31]. We show that EventDance performs well for the novel cross-modal (image-to-events) adaption task (see Fig. 2 (a)). We also show that it can be flexibly extended to a prior setting of *e.g.*, [68]: edge map to event image adaptation (see Fig.2 (b)). The experimental results demonstrate that our EventDance significantly outperforms the prior source-free domain adaptation methods *e.g.*, [33], in addressing the challenging cross-modal task. In summary, our main contributions are as follows: **(I)** We address a **novel yet challenging** problem for cross-modal (image-to-events) adaptation without access to the source image data. **(II)** We propose EventDance, which incorporates the RMB and MKA modules to fully exploit event. **(III)** Three event-based benchmarks with two adaptation settings demonstrate the effectiveness and superiority of EventDance.

2. Related Work

2.1. Event-based Object Recognition

aims to identify target objects from an event stream by taking full use of the event cameras' unique characteristics [74]. Since event cameras enjoy high temporal resolution, low latency, and very high dynamic range, this allows for real-time onboard object recognition in robotic, autonomous vehicles, and other mobile systems [70]. However, due to the distinct imaging paradigm shift, it is impossible to directly apply DNNs to learn events. Thus, various event representation types [5, 7, 8, 12–14, 22, 39, 59, 75] are proposed for mining the visual information and power from events, especially for the object recognition task. In previous works, *e.g.*, [68], diverse event representations are adopted as the target domain while failing to explore the raw events. *In this paper, we propose to learn target models that distinguish raw events and take multiple event representation for imposing consistency regularization.*

2.2. Cross-modal Knowledge Transfer

Knowledge transfer across modalities is first proposed in [23], aiming at learning representations for the modality with limited annotations based on a label-sufficient modal-

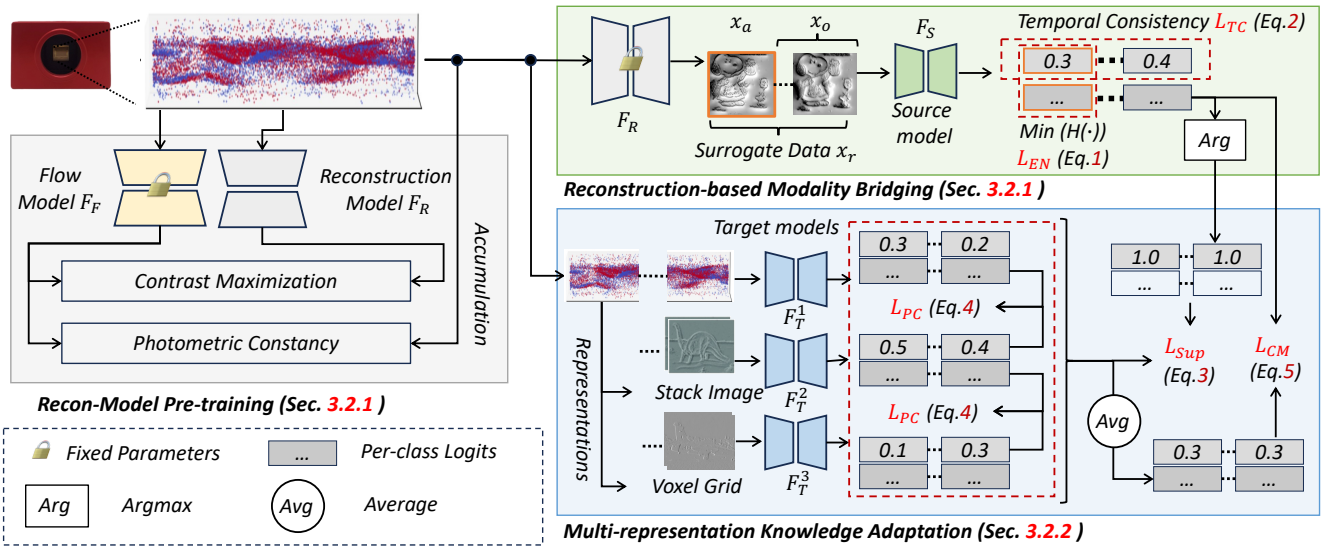


Figure 3. **Overall framework of our proposed framework**. RMB: reconstruction-based modality bridging module, MKA: multi-representation knowledge adaptation module.

ity. Increasing attention has been paid to the cross-modal knowledge transfer task for novel sensors, *e.g.*, event cameras. Most of the proposed methods [11, 25, 52] assume that the cross-modal paired data is achievable while some recent works tried to relax this assumption by reducing the required data [69]. Additionally, there are some works for classification based on domain translation [16, 18]. These methods rely on cross-modal data pairs and task-relevant paired data. To alleviate the demands on paired data, SOCKET [3] proposes an cross-modal adaptation framework that only utilizes external extra paired data for RGB-to-depth knowledge transfer, which are not always easy to obtain. *Differently, our EventDance is the* ***first*** *framework for cross-modal (image-to-event) adaptation without access to any source modality data, as shown in Fig. 2 (a). Due to the distinct modality gap with the image, we propose to build a surrogate domain and introducing representation consistency training for better knowledge transfer.*

2.3. Source-free UDA

UDA aims to alleviate the domain-shift problems caused by data distribution discrepancy in many computer vision tasks [26, 27, 30, 41, 45, 62, 64–66, 71–73]. However, the dependence on source data limits the generalization capability to some real applications, for reasons like data privacy issues [33]. Thus endeavors have been made in transferring knowledge only from the trained source models [1] without access to the source data. The cross-domain knowledge for unlabeled target data is extracted from single [36] or multiple [49] source models without access to the source data [33]. The ideas of source-free UDA can be formulated into two types according to whether the parameters of source models are available [17], *i.e.*, white-box and black-box models. Concretely, the white-box are achieved by data generation [15, 24, 51, 53] and model fine-tuning [9, 37] while the black-box depend on self-supervised learning [34, 61] and distribution alignment [60, 67].

CTN [68] is a UDA framework that leverages the edge maps obtained from the source RGB images and adapts the classification knowledge to a target model learning event images, as shown in Fig. 2 (b). In this paper, we focus on the source-free cross-modal (*i.e.*, image-to-event) adaptation without accessing the source data, which is essentially different from [68] and more challenging to tackle. *Our core idea is to create a surrogate domain in the image modality via the RMB module and update the surrogate domain for knowledge transfer via the MKA module.*

3. The Proposed Framework

3.1. Problem Setup and Overview

Knowledge adaptation from a source modality to a target modality can be more challenging than a domain shift between different datasets in the same modality, as demonstrated in [3]. Prior cross-modal adaption methods [3, 16, 18] predominantly depend on extra data from both modalities to bridge the source and target modalities. However, the novelty of event cameras, combined with the lack of this paired data, impedes the application of these techniques to event modality. Consequently, within the context of our cross-modal problem setup, we are limited to a pre-trained

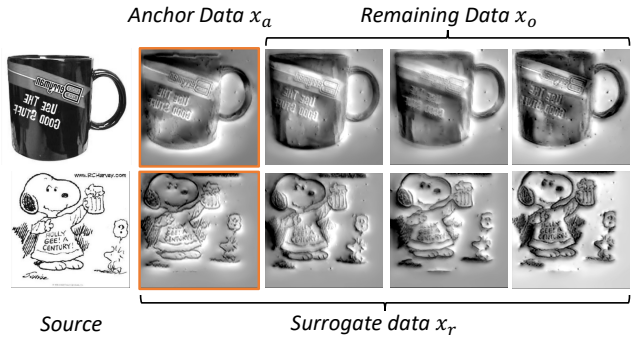


Figure 4. Visualization of samples in source and surrogate data in the image modality.

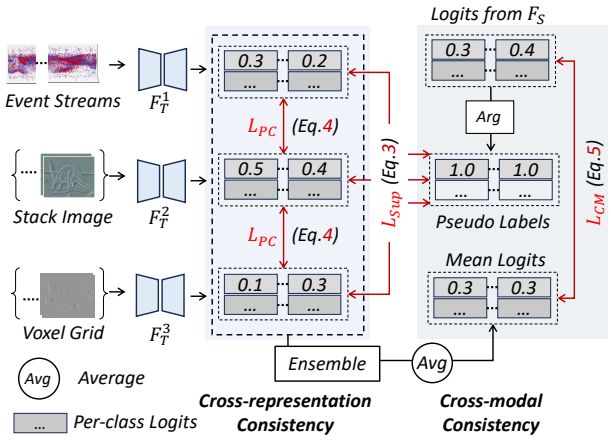


Figure 5. Illustration of the MKA module.

source model in the image modality and unlabeled target data in the event modality.

Our Key Idea: *By constructing the surrogate domain with target event data, we aim to mitigate modality gaps. This enables knowledge extraction from the source model. We subsequently employ multiple representations of the event data to accomplish the knowledge transfer.*

Primary objective: Denote the source model as F_S, where S indicates the source modality on which the model is trained. X_T represents the unlabeled target event data. As shown in Fig. 3, given a batch of event data $x_t \subset X_T$, we can obtain the surrogate image batch x_r using a reconstruction model F_R. Our aim is to derive target models F_T^i using both F_S and the unlabeled event data X_T. Here, the index i indicates the i-th target model, which ingests different event representation forms as input. Specifically, for $i = 1$, the input is a stack image; for $i = 2$, it is a voxel grid; and for $i = 3$, it is an event spike tensor (EST). In the following sections, we elaborate the proposed modules: reconstruction-based modality bridging (RMB) module (Sec. 3.2) and multi-representation knowledge adaptation (MKA) module (Sec. 3.3).

3.2. Reconstruction-based Modality Bridging

The RMB module builds a surrogate image domain to imitate the source image distribution. We utilize a self-supervised event-to-video model [44] to construct the **surrogate data in the image modality** directly from raw event data, as depicted in Fig. 4. The surrogate data facilitates the extraction of knowledge (*i.e.*, pseudo labels) from the image-trained source model.

However, simply introducing such a model is insufficient for our purpose. The reason is that it only focuses on generating natural-looking images and not optimal surrogate images that can effectively extract knowledge from the source model. Thus, we update the RMB module during training to generate better surrogate images for knowledge extraction. Specifically, we select the first surrogate image as the representative **anchor data** x_a, as shown in Fig. 4. Then, we pass it to the source model for predictions. Finally, we minimize the entropy of the prediction $F_S(x_a)$ to ensure that the surrogate images can effectively extract knowledge from the source model. We use $\mathcal{L}_{EN}$ to optimize the reconstruction model F_R, which can be formulated as:

$$\mathcal{L}_{EN} = min(H(F_S(x_a))), \tag{1}$$

where $H(\cdot)$ represents the entropy function.

As illustrated in Fig. 4, we also utilize the remaining data x_o (excluding the selected anchor data x_a) to augment the anchor data x_a and fully leverage the high-temporal resolution of events. To ensure temporal prediction consistency among the reconstructed images, we update F_S using the temporal consistency loss $\mathcal{L}_{TC}$:

$$\mathcal{L}_{TC} = \mathcal{L}_{kl}(F_S(x_a), F_S(x_o)), \tag{2}$$

where $\mathcal{L}_{kl}$ is the Kullback-Liibler (KL) divergence.

In practice, we adopt the framework proposed in [44] as our basic framework, which utilizes EvFlowNet [75] as the flow estimation model F_F and E2VID [50] as the reconstruction model F_R. EvFlowNet is trained with contrast maximization proxy loss [76] and provides accurate optical flow estimation. E2VID reconstructs intensity frames by exploring the flow-intensity relation with the event-based photometric constancy [19]. F_F and F_R are pre-trained using the unlabeled event data. Only F_R is updated during training with $\mathcal{L}_{EN}$, while F_F remains fixed to prevent F_R from the model collapse. With the knowledge extraction module, we obtain the prediction logits and pseudo labels P from $F_S(x_a)$ for learning the target event-based models.

3.3. Multi-representation Knowledge Adaptation

Though our RMB module facilitates mitigating the modality gaps, adapting knowledge from images to events remains challenging due to: **1)** a single event representation

type, *e.g.*, voxel grid [68], cannot comprehensively represent event data, leading to information loss during the adaptation process, and **2)** the source model is not ideal for cross-modal knowledge transfer, hindering the transfer efficiency. To this end, we propose to learn multiple target models using distinct event representations to fully leverage the high temporal resolution of events, as shown in Fig. 3.

To process raw event data, we convert a given event stream E into commonly used event representations. For voxel grid, as proposed in [63], we obtain the voxel grid $E_v \subset \mathbb{R}^{H\times W\times C}$ with B temporal bins using consecutive and non-overlapping segments of E, where H, W, and C are spatial sizes. E_v adaptively normalizes the temporal dimension of the input based on the timestamps of each segment of the event stream. For the event stack images, we employ a stacking strategy [32] to sample and stack events in a fixed constant number. These results in a tensor-like representation $E_s \subset \mathbb{R}^{H\times W\times 1}$. For EST, we directly use the method in [21] taking raw events as input.

Technically, as shown in Fig. 5, we set two training objectives for the target models as: **1)** *cross-representation consistency* training with several event representation types in training target models F_T^i and **2)** *cross-modal consistency* training between the source model F_S and i-th target model F_T^i. Based on the RMB module, the pseudo labels P are obtained through the $argmax$ operation applied to $F_S(x_a)$. We denote the i-th event representation for F_T^i as $r(x_t)^i$. The target models are supervised by the pseudo labels P with the cross entropy $\mathcal{L}_{ce}$ as:

$$\mathcal{L}_{Sup} = \mathcal{L}_{ce}(F_S^i(r(x_t)^i), P), i \in \{1, 2, 3\}. \quad (3)$$

$\mathcal{L}_{Sup}$ serves as the fundamental knowledge transfer loss.

Cross-representation consistency. These three event representation types are fed into their corresponding target models, and prediction consistency training is conducted among the models to facilitate target model learning from each other. The prediction consistency training loss among different event representation types can be formulated as:

$$\mathcal{L}_{PC} = \sum_{k=1,l\neq k}^{3} \{\mathcal{L}_{kl}(F_T^k(r(x_t)^k), F_T^l(r(x_t)^l))\}. \quad (4)$$

Cross-modal consistency. Additionally, as the source model is not an ideal model for image-to-event transfer, we propose a cross-modal consistency learning strategy between the source model F_S and target model F_T^i to simultaneously update both models. This improves the performance of F_S and makes it more suitable for image-to-event knowledge transfer. The average ensemble results from F_T^i and F_S are mutually supervised by each other. The cross-modal consistency loss is formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{CM}} =& \mathcal{L}_{kl}(\frac{1}{3}\sum_{i=1}^{3}(F_T^i(r(x_t)^i)), F_S(x_a)) \\ &+\mathcal{L}_{kl}(F_S(x_a), \frac{1}{3}\sum_{i=1}^{3}(F_T^i(r(x_t)^i))).\end{aligned} \quad (5)$$

Overall, our final loss is the combination of the above losses in Eq. 1, 2, 3, 4, and 5. The overall loss function $\mathcal{L}_{all}$:

$$\mathcal{L}_{all} = \mathcal{L}_{EN} + \mathcal{L}_{TC} + \mathcal{L}_{Sup} + \mathcal{L}_{PC} + \mathcal{L}_{CM}. \quad (6)$$

Concretely, $\mathcal{L}_{EN}$ is used to optimize F_R; $\mathcal{L}_{TC}$ and $\mathcal{L}_{CM}$ are used to optimize F_S; and $\mathcal{L}_{Sup}$, $\mathcal{L}_{PC}$, and $\mathcal{L}_{CM}$ are used to optimize F_T^i. The whole framework is optimized in an end-to-end manner.

4. Experiments

In this section, we empirically validate various aspects of EventDance. In Sec. 4.1, we show the experimental settings of our image-to-event adaptation, baseline, comparison methods, and implementation details. We further show the performance of EventDance compared with the existing cross-modal and UDA methods in Sec. 4.2.

4.1. Datasets and Implementation Details

N-MNIST[43] is the event-based version of the well-known MNIST dataset. The dataset was created by capturing the visual input of an event camera focused on a monitor displaying the original MNIST data. **N-CALTECH101**[43] is the event-based extension of the CALTECH101 dataset, which contains 100 object classes along with a background class. This dataset poses a challenge due to the many classes and the unbalanced number of samples within each class. **CIFAR10-DVS** [31] consists of 10,000 event streams belonging to 10 classes, captured by a DVS camera with a spatial resolution of 128 × 128.

Evaluation configurations. EventDance employs three target models by taking three event representation types (stack image, voxel grid, and EST) as inputs in the training phase, respectively. Therefore, *the inference can be achieved using one of the models*. We present the recognition accuracy results of the target models taking voxel gird, on three event-based benchmarks in Tab. 1.

Baseline and comparison methods. As we are the first to address the cross-modal problem, there is no direct baseline available for comparison. We establish a baseline in Tab. 6 and Tab. 1 to evaluate the performance of the pre-trained source model with event voxel grids, as in the previous work [68]. Also, we compare our method with the existing SFUDA methods [33, 35], image-to-voxel-grid adaptation method [68], and the UDA method using source data [77] that use event voxel grids as the target modality data.

Implementation Details. We use ResNet-18, 34, and 50

Method	Backbone	N-MNIST	Δ	CIFAR10-DVS	Δ	N-CALTECH101	Δ
Baseline	R-18	41.80	-	36.14	-	40.81	-
	R-34	69.10	-	45.29	-	58.78	-
	R-50	77.10	-	60.88	-	70.73	-
SHOT [33]	R-18	53.70	+11.90	36.97	+0.84	44.26	+3.45
	R-34	78.40	+9.30	46.32	+1.03	61.12	+2.34
	R-50	88.90	+11.80	61.03	+0.15	83.35	+12.62
Zhao *et al*. [68]	R-18	53.20	+11.40	36.42	+0.28	44.30	+3.49
	R-34	76.90	+7.80	45.78	+0.49	61.10	+2.32
	R-50	84.60	+7.50	61.99	+1.11	78.72	+7.99
SHOT++ [35]	R-18	68.80	<u>+27.00</u>	37.26	+1.12	49.33	+8.52
	R-34	84.70	+15.60	46.37	+1.08	64.54	+5.76
	R-50	<u>89.40</u>	+12.30	63.41	+2.53	<u>82.88</u>	+12.15
EventDance (Ours)	R-18	71.00	**+29.20**	62.13	<u>+25.99</u>	66.77	**+25.96**
	R-34	86.50	+17.40	<u>71.98</u>	**+26.69**	72.68	+13.90
	R-50	**92.30**	+15.20	**85.69**	+24.81	**92.35**	<u>+21.62</u>

Table 1. Experimental results on images-to-events with **SFUDA** methods (see Fig. 2 (a)). Δ: The performance gain over the baseline. The **bold** and <u>underline</u> denote the best and the second-best performance in SFUDA methods, respectively.

Method	S.F.	Unsup.	Backbone / Train	N-CAL
E2VID [50]	✗	✓	Fine-tune	59.80
+ CLIP	✗	✓	Scratch	9.40
Ev-LaFOR [10]	✗	✓	Text Prompt	82.46
+ CLIP	✗	✓	Visual Prompt	82.61
Wang *et al*. [56]	✓	✓	-	42.70
	✗	✓	-	43.50
+ CLIP	✓	✗	-	39.70
DSAN [77]	✗	✓	R-18	78.45
	✗	✓	R-34	89.01
	✗	✓	R-50	94.56
EventDance (Ours)	✓	✓	R-18	66.77
	✓	✓	R-34	72.68
	✓	✓	R-50	92.35

Table 2. Experimental results compared with label-free methods.

Method	S.F.	Unsup.	Backbone	N-MNIST
EV-VGCNN [14]	✗	✗	EV-VGCNN	99.10
Deep SNN [29]	✗	✗	Deep SNN	98.70
Phased LSTM [42]	✗	✗	Phased LSTM	97.30
PointNet++ [58]	✗	✗	PointNet++	95.50
EventDance (Ours)	✓	✓	R-18	71.00
	✓	✓	R-34	86.50
	✓	✓	R-50	92.30

Table 3. Experimental results compared with supervised methods.

(R-18, R-34, and R-50) pre-trained on ImageNet as backbones. The batch size is set to 64, following the prior work [68]. We use the AdamW optimizer with a learning rate of 1e-5, which linearly decays over time. We use

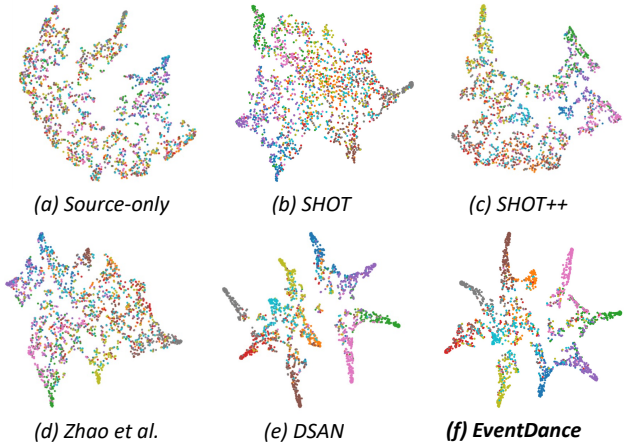
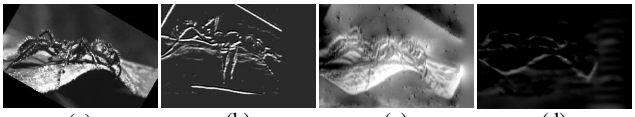


Figure 6. TSNE [54] visualization of (a) source-only, (b) SHOT, (c) SHOT++, (d) Zhao *et al*., (e)DSAN, and (f) EventDance on the target modality CIFAR10-DVS dataset with R-18 backbone. Different colors represent the 10 classes in CIFAR10-DVS dataset.

image augmentation techniques, *e.g*., random rotations and flipping for source modality pre-training. However, we do not use event augmentation techniques during target learning for a fair comparison with other methods. *More details about the settings can be found in the supplmat.*

4.2. Experimental Results

We evaluate our EventDance under the challenging source-free image-to-events adaptation setting. The experimental results are shown in Tab. 1. EventDance consistently outperforms the source-free UDA methods [33, 35], source-free cross-modal UDA method [68] and even achieves recognition accuracy closer to that of the UDA method



(a) (b) (c) (d)

Figure 7. Visualization of (a) source gray-scale image; (b) event stack image; (c) reconstructed image before training; (d) reconstructed image after training.

DSAN [77] that utilizes the source data on the three event-based benchmarks. EventDance brings significant performance gains of +25.99%, +26.69%, and +24.81% with R-18, R-34, and R-50 backbones, respectively. This indicates the superiority of our proposed RMB and MKA modules in tackling the non-trivial cross-modal problem.

As shown in Tab. 2, we also compare our EventDance with the label-free methods, such as E2VID [50] + CLIP, Ev-LaFOR [10] + CLIP, Wang *et al*. [56] + CLIP, and UDA method DSAN [77]. Obviously, our EventDance outperforms these label-free methods and achieves recognition accuracy that is closer to that of the UDA method DSAN (with source data) [77], even without using source modality data (92.35% vs. 94.56% with R-50 backbone).

Furthermore, we provide a performance comparison between our EventDance and several state-of-the-art supervised event-based recognition methods on the N-MNIST dataset in Tab. 3, including EV-VGCNN [14], Deep SNN [29], Phased LSTM [42], and PointNet++ [58]. Our EventDance achieves good performance in an unsupervised manner, without using the source data. We provide the TSNE [54] visualization in Fig. 6, apparently, our EventDance brings a significant improvement in distinguishing cross-modal samples in high-level feature space.

5. Ablation Study and Analysis

Different combination of proposed modules. To validate the effectiveness of the proposed modules, we conduct experiments on the NCALTECH-101 dataset with different combinations of modules. Tab. 4 shows the detailed results of the performance with different loss and component combinations. All of our proposed modules and loss functions have a positive impact on improving recognition accuracy. Notably, fine-tuning the reconstruction model F_R results in a significant performance gain by 17.83%, which supports our claim of building a surrogate data in the image modality for better knowledge transfer, rather than the visual quality of event-to-video reconstruction. This is further supported by the visual results presented in Fig. 7.

Event representation vs. target model's performance. For a fair comparison, we validate the quantitative results of all methods in Tab. 6 and Tab. 1 using event voxel grids. To investigate how to fully leverage the abundant spatio-temporal information of events for object recognition, we provide the results of validating our EventDance with different representation types in Tab. 5. Compared to inference with voxel grids, using EST, which contains more spatio-temporal information of events, achieves the best recognition accuracy gains by +0.19%, +0.52%, and +0.39%, on the backbones R-18, R-34, and R-50, respectively. The results reflect that EST is better suited for object recognition. The stack images, for which temporal information is lost, achieve lower recognition accuracy than the voxel grids by -0.07%, -1.52%, and -0.81% with backbones R-18, R-34, and R-50, respectively. Thus, it is crucial to explore the event data with representations that remain temporal characteristic of events in the cross-modal adaptation problems.

F_R	$\mathcal{L}_{TC}$	$\mathcal{L}_{EN}$	$\mathcal{L}_{Sup}$	$\mathcal{L}_{PC}$	$\mathcal{L}_{CM}$	Accurracy	Δ
			✓			40.81	-
		✓	✓			43.36	+2.55
	✓		✓			45.89	+5.08
✓			✓			58.64	+17.83
	✓	✓	✓			53.40	+10.04
✓	✓	✓	✓			61.83	+21.02
			✓		✓	43.38	+2.57
			✓	✓		42.56	+1.75
✓	✓	✓	✓		✓	63.58	+22.77
✓	✓	✓	✓	✓		63.26	+22.45
✓	✓	✓	✓	✓	✓	66.77	+25.96

Table 4. Ablation study of different module combinations on N-CALTECH101 with ResNet-18.

Backbone	Event Representations		
	Stack Image	Voxel Grid	Event Spike Tensor
R-18	$66.70_{-0.07}$	66.77	$66.96_{+0.19}$
R-34	$71.16_{-1.52}$	72.68	$73.00_{+0.52}$
R-50	$91.54_{-0.81}$	92.35	$92.74_{+0.39}$

Table 5. Ablation experiments on the inference of our proposed method with different event representations.

Ablation of RMB module. The RMB module is a crucial component for bridging the modality gaps between images and events. The results are shown in Tab. 4, where we use a checkmark to denote F_R fine-tuning. As can be seen, only using the pre-trained reconstruction model F_R to construct the surrogate data achieves the recognition accuracy of 53.04%. Obviously, updating F_R improves the accuracy to 61.83%. We visualize the surrogate data's intensity frames, as shown in Fig. 7. Although the reconstructed image in Fig. 7 (d) after training is less distinct than the one in Fig. 7 (c) before training, the ability to extract knowledge becomes significantly better (53.04% vs. 61.83%).

Method	Source-Free	Unsupervised	Backbone	N-MNIST	Δ	CIFAR10-DVS	Δ	N-CALTECH101	Δ
Baseline	✗	✓	R-18	82.60	-	54.10	-	72.80	-
			R-34	84.30	-	55.70	-	73.20	-
			R-50	84.70	-	56.50	-	74.70	-
Zhao *et al*. [68]	✓	✓	R-18	98.60	+16.00	76.50	+22.40	88.50	+15.70
			R-34	99.00	+14.70	76.70	+21.00	89.30	+16.10
			R-50	99.30	+14.60	77.30	+20.80	90.10	+15.40
EventDance (ours)	✓	✓	R-18	99.10	**+16.50**	79.80	+25.70	90.30	+17.50
			R-34	<u>99.40</u>	<u>+15.10</u>	<u>85.40</u>	<u>+29.70</u>	<u>91.40</u>	**+18.20**
			R-50	**99.70**	+15.00	**86.70**	**+30.20**	**92.30**	<u>+17.60</u>

Table 6. Experimental results on edge maps-to-voxel grids with UDA methods (see Fig. 2 (a)). Test Rep.: event representation used in test; VG: voxel grid. The **bold** and <u>underline</u> denote the best and the second-best performance in source-free uda methods, respectively.

Representation	S	V	E	All
Accuracy	61.63	63.58	65.74	**66.91**

Table 7. Ablation on the usage of event representations in target model training. (S: stack image; V: voxel grid; E: EST.)

6. Extension Experiment

We show that our method can be flexibly extended to the adaptation problem from edge maps to event voxel grids, as done in [68]. Specifically, we test our EventDance under the source-free UDA setting [68], where event streams are converted into event voxel grids with $C = 3$ as the target data, and images are processed to edge maps as source modality data for pre-training source models. As in Tab. 6, our approach consistently outperforms the SoTA method CTN [68] on three event-based benchmarks.

7. Discussion

Performance gains in two experimental settings. Edge maps have a closer similarity to voxel grids of events, which makes the knowledge transfer easier compared to our cross-modal setting. Therefore, our EventDance achieves a greater performance gain in the image-to-events setting (26.15% w/ R-18) compared to the edge maps-to-voxel grids (+17.50% w/ R-18) on NCALTECH-101.

High temporal resolution of events. Cross-modal knowledge transfer is challenging due to the distinct modality gap between images and events, namely $H \times W \times C$ for images and (x, y, t, p) for events [70]. The straightforward approach to alleviating this problem is to convert events into image-like tensors. However, most event representations struggle with information loss, such as temporal information. For downstream tasks, there might be the best event representation that achieves the SoTA performance, such as EST [21] in object recognition. Nevertheless, we find that multiple event representations are suitable for source-free cross-modal adaptation. This observation is supported by the quantitative results shown in Tab. 7.

Surrogate data in training. While building surrogate data incurs higher computation costs, it effectively eliminates the need for extra paired data that may not always be available in practice. Moreover, in our work, the reconstruction model used to construct the surrogate data is only trained and updated during the training phase and can be freely discarded during the inference.

Selection of anchor data. We experimentally determine to select the first surrogate image as anchor data, which is reconstructed from the initial period of the event stream. This is the most effective method to obtain reliable anchor data, as the size of event streams varies across the target dataset. The remaining frames are treated as augmentation for anchor data. Attempts at random selection result in low-quality images for shorter streams.

8. Conclusion

In this paper, we investigated a new problem of achieving image-to-event adaptation for event-based object recognition without access to any source image. To this end, we proposed an cross-modal framework, named EventDance. The experiments for image-to-events and edge maps-to-voxel grids adaptation show that EventDance outperforms prior source-free and cross-modal UDA methods and is on par with the methods that use source data.

Limitation and Future Work: One limitation of EventDance is that training three target models with different representations lead to increased computational costs during training. However, our method has significant implications for the event-based vision and may open a new research direction. In the future, we plan to extend our approach to other downstream tasks.

Acknowledgement This paper is supported by the National Natural Science Foundation of China (NSF) under Grant No. NSFC22FYT45 and the Guangzhou City, University and Enterprise Joint Fund under Grant No.SL2022A03J01278.

References

[1] Sk Miraj Ahmed, Aske R. Lejbølle, Rameswar Panda, and Amit K. Roy-Chowdhury. Camera on-boarding for person re-identification using hypothesis transfer learning. In *CVPR*, pages 12141–12150. Computer Vision Foundation / IEEE, 2020. 2, 3

[2] Sk Miraj Ahmed, Dripta S Raychaudhuri, Sujoy Paul, Samet Oymak, and Amit K Roy-Chowdhury. Unsupervised multi-source domain adaptation without access to source data. In *CVPR*, pages 10103–10112, 2021. 2

[3] Sk Miraj Ahmed, Suhas Lohit, Kuan-Chuan Peng, Michael Jones, and Amit K. Roy-Chowdhury. Cross-modal knowledge transfer without task-relevant source data. In *ECCV*, pages 111–127. Springer, 2022. 2, 3

[4] Himanshu Akolkar, Sio-Hoi Ieng, and Ryad Benosman. Real-time high speed motion prediction using fast aperture-robust event-driven visual flow. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):361–372, 2022. 1

[5] Mohammed Almatrafi, Raymond Baldwin, Kiyoharu Aizawa, and Keigo Hirakawa. Distance surface for event-based optical flow. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1547–1556, 2020. 2

[6] Raymond Baldwin, Ruixu Liu, Mohammed Mutlaq Almatrafi, Vijayan K Asari, and Keigo Hirakawa. Time-ordered recent event (tore) volumes for event cameras. *IEEE TPAMI*, 2022. 1

[7] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 136–152. Springer, 2020. 2

[8] Jiahang Cao, Xu Zheng, Yuanhuiyi Lyu, Jiaxu Wang, Renjing Xu, and Lin Wang. Chasing day and night: Towards robust and efficient all-day object detection guided by an event camera. *arXiv preprint arXiv:2309.09297*, 2023. 2

[9] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7354–7362. Computer Vision Foundation / IEEE, 2019. 3

[10] Hoonhee Cho, Hyeonseong Kim, Yujeong Chae, and Kuk-Jin Yoon. Label-free event-based object recognition via joint learning with image reconstruction from events. *arXiv preprint arXiv:2308.09383*, 2023. 6, 7

[11] Rui Dai, Srijan Das, and François Brémond. Learning an augmented RGB representation with cross-modal knowledge distillation for action detection. In *ICCV*, pages 13033–13044. IEEE, 2021. 3

[12] Yongjian Deng, Youfu Li, and Hao Chen. Amae: Adaptive motion-agnostic encoder for event-based object classification. *IEEE Robotics and Automation Letters*, 5(3):4596–4603, 2020. 2

[13] Yongjian Deng, Hao Chen, and Youfu Li. Mvf-net: A multi-view fusion network for event-based object classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8275–8284, 2021.

[14] Yongjian Deng, Hao Chen, Hai Liu, and Youfu Li. A voxel graph cnn for object classification with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1172–1181, 2022. 2, 6, 7

[15] Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *CVPR*, pages 7202–7212. IEEE, 2022. 3

[16] Dapeng Du, Limin Wang, Huiling Wang, Kai Zhao, and Gangshan Wu. Translate-to-recognize networks for RGB-D scene recognition. In *CVPR*, pages 11836–11845. Computer Vision Foundation / IEEE, 2019. 3

[17] Yuqi Fang, Pew-Thian Yap, Weili Lin, Hongtu Zhu, and Mingxia Liu. Source-free unsupervised domain adaptation: A survey. *CoRR*, abs/2301.00265, 2023. 3

[18] Andrea Ferreri, Silvia Bucci, and Tatiana Tommasi. Translate to adapt: RGB-D scene recognition across domains. *CoRR*, abs/2103.14672, 2021. 3

[19] Guillermo Gallego, Christian Forster, Elias Mueggler, and Davide Scaramuzza. Event-based camera pose tracking using a generative event model. *arXiv preprint arXiv:1510.01972*, 2015. 4

[20] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):154–180, 2022. 1

[21] 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, 5, 8

[22] 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/CVF International Conference on Computer Vision*, pages 5633–5643, 2019. 2

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

[24] Yan Hao, Yuhong Guo, and Chunsheng Yang. Source-free unsupervised domain adaptation with surrogate data generation. In *BMVC*, page 198. BMVA Press, 2021. 3

[25] Judy Hoffman, Saurabh Gupta, Jian Leong, Sergio Guadarrama, and Trevor Darrell. Cross-modal adaptation for RGB-D detection. In *ICRA*, pages 5032–5039. IEEE, 2016. 3

[26] Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao. Category contrast for unsupervised domain adaptation in visual tasks. In *CVPR*, pages 1193–1204. IEEE, 2022. 3

[27] Zhipeng Huang, Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Peng Chu, Quanzeng You, Jiang Wang, Zicheng Liu, and Zheng-Jun Zha. Lifelong unsupervised domain adaptive person re-identification with coordinated anti-forgetting and adaptation. In *CVPR*, pages 14268–14277. IEEE, 2022. 3

[28] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE TPAMI*, 39(7):1346–1359, 2016. 1

[29] Jun Haeng Lee, Tobi Delbruck, and Michael Pfeiffer. Training deep spiking neural networks using backpropagation. *Frontiers in neuroscience*, 10:508, 2016. 6, 7

[30] Taeyeop Lee, Byeong-Uk Lee, Inkyu Shin, Jaesung Choe, Ukcheol Shin, In So Kweon, and Kuk-Jin Yoon. UDA-COPE: unsupervised domain adaptation for category-level object pose estimation. In *CVPR*, pages 14871–14880. IEEE, 2022. 3

[31] 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, 5

[32] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Trans. Image Process.*, 31:2975–2987, 2022. 2, 5

[33] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, pages 6028–6039. PMLR, 2020. 2, 3, 5, 6

[34] Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. DINE: domain adaptation from single and multiple black-box predictors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7993–8003. IEEE, 2022. 3

[35] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):8602–8617, 2022. 5, 6

[36] Qing Liu, Adam Kortylewski, Zhishuai Zhang, Zizhang Li, Mengqi Guo, Qihao Liu, Xiaoding Yuan, Jiteng Mu, Weichao Qiu, and Alan L. Yuille. Learning part segmentation through unsupervised domain adaptation from synthetic vehicles. In *CVPR*, pages 19118–19129. IEEE, 2022. 3

[37] Xinyu Liu and Yixuan Yuan. A source-free domain adaptive polyp detection framework with style diversification flow. *IEEE Trans. Medical Imaging*, 41(7):1897–1908, 2022. 3

[38] Jacques Manderscheid, Amos Sironi, Nicolas Bourdis, Davide Migliore, and Vincent Lepetit. Speed invariant time surface for learning to detect corner points with event-based cameras. In *CVPR*, pages 10245–10254, 2019. 1

[39] 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. 2

[40] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters*, 7(2):3515–3522, 2022. 1

[41] Muhammad Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *CVPR*, pages 14745–14755. IEEE, 2022. 3

[42] Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. Phased lstm: Accelerating recurrent network training for long or event-based sequences. *Advances in neural information processing systems*, 29, 2016. 6, 7

[43] Garrick Orchard, Ajinkya Jayawant, Gregory Cohen, and Nitish V. Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *CoRR*, abs/1507.07629, 2015. 2, 5

[44] Federico Paredes-Vallés and Guido CHE de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021. 2, 4

[45] Sujoy Paul, Yi-Hsuan Tsai, Samuel Schulter, Amit K. Roy-Chowdhury, and Manmohan Chandraker. Domain adaptive semantic segmentation using weak labels. In *ECCV*, pages 571–587. Springer, 2020. 3

[46] Michaël Perrot and Amaury Habrard. A theoretical analysis of metric hypothesis transfer learning. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1708–1717. JMLR.org, 2015. 2

[47] Christoph Posch, Teresa Serrano-Gotarredona, Bernabé Linares-Barranco, and Tobi Delbrück. Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output. *Proc. IEEE*, 102(10):1470–1484, 2014. 1

[48] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43 (6):1964–1980, 2021. 1

[49] Chuan-Xian Ren, Yong Hui Liu, Xiwen Zhang, and Ke-Kun Huang. Multi-source unsupervised domain adaptation via pseudo target domain. *IEEE TIP*, 31:2122–2135, 2022. 3

[50] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 156–163, 2020. 4, 6, 7

[51] Serban Stan and Mohammad Rostami. Unsupervised model adaptation for continual semantic segmentation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2593–2601. AAAI Press, 2021. 3

[52] Fida Mohammad Thoker and Juergen Gall. Cross-modal knowledge distillation for action recognition. In *2019 IEEE ICIP, Taipei, Taiwan, September 22-25, 2019*, pages 6–10. IEEE, 2019. 3

[53] Jiayi Tian, Jing Zhang, Wen Li, and Dong Xu. VDM-DA: virtual domain modeling for source data-free domain adaptation. *IEEE Trans. Circuits Syst. Video Technol.*, 32(6):3749–3760, 2022. 3

[54] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 6, 7

[55] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Ultimate slam? combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios. *IEEE Robotics Autom. Lett.*, 3(2):994–1001, 2018. 1

[56] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10081–10090, 2019. 6, 7

[57] Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 608–619, 2021. 1

[58] Qinyi Wang, Yexin Zhang, Junsong Yuan, and Yilong Lu. Space-time event clouds for gesture recognition: From rgb cameras to event cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1826–1835. IEEE, 2019. 6, 7

[59] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 6358–6367, 2019. 2

[60] Baoyao Yang, Hao-Wei Yeh, Tatsuya Harada, and Pong C. Yuen. Model-induced generalization error bound for information-theoretic representation learning in source-data-free unsupervised domain adaptation. *IEEE Trans. Image Process.*, 31:419–432, 2022. 3

[61] Taojiannan Yang, Sijie Zhu, Chen Chen, Shen Yan, Mi Zhang, and Andrew R. Willis. Mutualnet: Adaptive convnet via mutual learning from network width and resolution. In *ECCV*, pages 299–315. Springer, 2020. 3

[62] Xu Yang, Cheng Deng, Tongliang Liu, and Dacheng Tao. Heterogeneous graph attention network for unsupervised multiple-target domain adaptation. *IEEE TPAMI*, 44(4): 1992–2003, 2022. 3

[63] Chengxi Ye, Anton Mitrokhin, Cornelia Fermüller, James A. Yorke, and Yiannis Aloimonos. Unsupervised learning of dense optical flow, depth and egomotion with event-based sensors. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2020, Las Vegas, NV, USA, October 24, 2020 - January 24, 2021*, pages 5831–5838. IEEE, 2020. 2, 5

[64] Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking. In *CVPR*, pages 8886–8895. IEEE, 2022. 3

[65] Chunyan Yu, Caiyu Liu, Meiping Song, and Chein-I Chang. Unsupervised domain adaptation with content-wise alignment for hyperspectral imagery classification. *IEEE Geosci. Remote. Sens. Lett.*, 19:1–5, 2022.

[66] Jingyi Zhang, Jiaxing Huang, Zichen Tian, and Shijian Lu. Spectral unsupervised domain adaptation for visual recognition. In *CVPR*, pages 9819–9830. IEEE, 2022. 3

[67] Wen Zhang and Dongrui Wu. Discriminative joint probability maximum mean discrepancy (DJP-MMD) for domain adaptation. In *IJCNN*, pages 1–8. IEEE, 2020. 3

[68] Junwei Zhao, Shiliang Zhang, and Tiejun Huang. Transformer-based domain adaptation for event data classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 4673–4677. IEEE, 2022. 1, 2, 3, 5, 6, 8

[69] 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 6527–6536. Computer Vision Foundation / IEEE, 2020. 3

[70] Xu Zheng, Yexin Liu, Yunfan Lu, Tongyan Hua, Tianbo Pan, Weiming Zhang, Dacheng Tao, and Lin Wang. Deep learning for event-based vision: A comprehensive survey and benchmarks, 2023. 1, 2, 8

[71] Xu Zheng, Tianbo Pan, Yunhao Luo, and Lin Wang. Look at the neighbor: Distortion-aware unsupervised domain adaptation for panoramic semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18687–18698, 2023. 3

[72] Xu Zheng, Jinjing Zhu, Yexin Liu, Zidong Cao, Chong Fu, and Lin Wang. Both style and distortion matter: Dual-path unsupervised domain adaptation for panoramic semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1285–1295, 2023.

[73] Xu Zheng, Pengyuan Zhou, Athanasios Vasilakos, and Lin Wang. Semantics, distortion, and style matter: Towards source-free uda for panoramic segmentation, 2024. 3

[74] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Eventbind: Learning a unified representation to bind them all for event-based open-world understanding, 2024. 2

[75] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018. 2, 4

[76] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 4

[77] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. Deep subdomain adaptation network for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4):1713–1722, 2021. 5, 6, 7

[78] Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, and Ying Fu. Learning to reconstruct high speed and high dynamic range videos from events. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2024–2033. Computer Vision Foundation / IEEE, 2021. 1