Bridging the Gap between Events and Frames through Unsupervised Domain Adaptation

Nico Messikommer, Daniel Gehrig, Mathias Gehrig, Davide Scaramuzza
Dept. of Informatics, Univ. of Zurich and Dept. of Neuroinformatics, Univ. of Zurich and ETH Zurich

Abstract

Event cameras are novel sensors with outstanding properties such as high temporal resolution and high dynamic range. However, event-based vision has been held back by the shortage of labeled datasets due to the novelty of event cameras. To overcome this drawback, we propose a task transfer method to train models directly with labeled images and unlabeled event data. We leverage the generative event model to split event features into content and motion features. Thus, our approach unlocks the vast amount of existing image datasets for the training of event-based neural networks. Our task transfer method outperforms methods targeting Unsupervised Domain Adaptation for object detection by 0.26 mAP and classification by 2.7% accuracy.

1. Introduction

The outstanding properties such as high dynamic range, high temporal resolution, and low latency make event cameras promising for several computer vision applications in edge-case scenarios. However, event cameras suffer from the scarcity of labeled datasets since event-based datasets represent only 3.14% of the existing vision dataset [2, 8].

Instead of capturing images at a fixed rate, event cameras measure changes in intensity asynchronously per pixel. This results in a stream of events that encodes the time, location, and polarity of the intensity change. For a more in-depth survey, we refer to [3]. Despite the radical different working principle, the output of event and frame-based cameras still contains a significant information overlap, as both cameras share the underlying principle of capturing the scene irradiance through an optical system [14]. In this work, we show how this information overlap can be leveraged for *Unsupervised Domain Adaptation (UDA)* of event-based networks, in which labeled source (image domain Y_{img}) and unlabeled target data (event domain Y_{event}) are available to transfer a task to the target domain. Code can be found at `https://github.com/uzh-rpg/rpg_ev-transfer` and additional results at `https://youtu.be/fZnBSqni6PY`

2. Method

In our framework, events $\mathbf{y}_{\text{event}}$ and images $\mathbf{y}_{\text{img}}$ are processed with separate encoders E_{img} and E_{event} due to the large domain gap between Y_{img} and Y_{event}, as shown in Fig. 1. As the asynchronous output signal of event cameras also contains motion information, event cameras measure specific features ζ_{event} about the scene, which standard cameras can not perceive in a single frame. This non-overlapping information, however, hinders the image-to-event task transfer as it is impossible to fully align the embedding space. We solve this by separating event features into *sensor specific* features ζ_{event} computed by a specific encoder $E_{\text{event, attr}}$, and *content* features $\mathbf{z}_{\text{event}}$, which contain information shared in both domains Y_{img} and Y_{event}.

$$\begin{aligned} \mathbf{z}_{\text{img}} &= E_{\text{img}}(\mathbf{y}_{\text{img}}) \\ \mathbf{z}_{\text{event}} = E_{\text{event}}(\mathbf{y}_{\text{event}}) \qquad \zeta_{\text{event}} &= E_{\text{event, attr}}(\mathbf{y}_{\text{event}}). \end{aligned} \tag{1}$$

The resulting shared features $\mathbf{z}_{\text{img}}$ and $\mathbf{z}_{\text{event}}$ are given as input to the task branch T, which computes the task-specific output. To generate pseudo event and image pairs, shared features from an image $\mathbf{z}_{\text{img}}$ are combined with event-specific features ζ_{event} from a random event sample to compute a pseudo-flow field using a flow decoder D_{flow}. The resulting pseudo-flow and the input image are then converted to events $\hat{\mathbf{y}}_{\text{event}}$ in the refinement network R_{ref}, The overall architecture is depicted in Fig. 1.

To enforce the embedding alignment, we apply adversarial training [7] with a PatchGAN discriminator network F_{lat} [9] to the latent features $\mathbf{z}_{\text{img}}$ and $\mathbf{z}_{\text{event}}$ and introduce L^1 consistency losses $\mathcal{L}_{\text{cycle}}$ on the latent variables $\mathbf{z}_{\text{img}}$ and ζ_{event}. The generation of realistic events from a single image is enforced by additional adversarial losses $\mathcal{L}_{\text{recons.,disc.}}$ and $\mathcal{L}_{\text{recons.,gen.}}$, which are applied on the reconstructed events $\hat{\mathbf{y}}_{\text{event}}$ using an event discriminator F_{event}. Finally, the task loss $\mathcal{L}_{task}$ is applied on the images and the fake events, which both have corresponding image labels. The used constraints are visualized with red arrows in Fig. 1

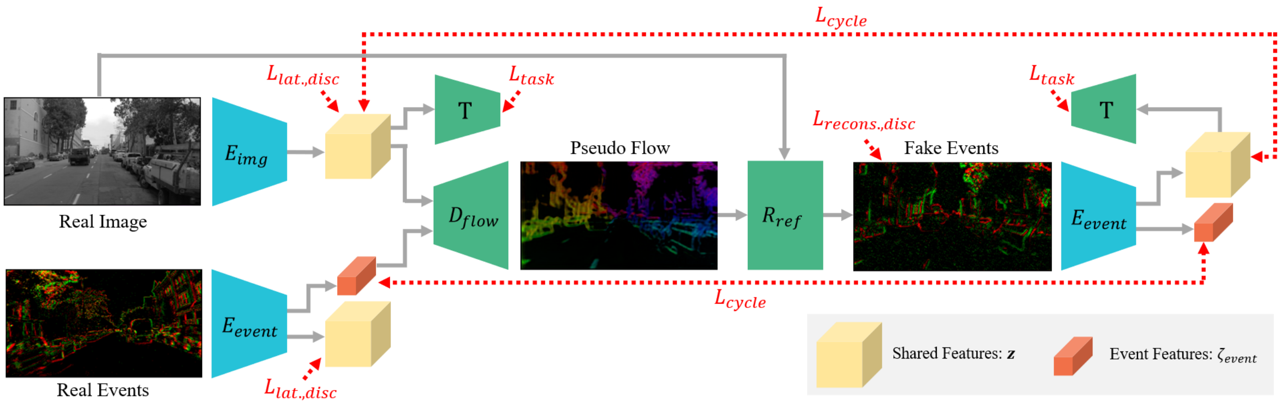

Figure 1. As there is a large domain gap between events and grayscale images, we use two separate encoders E_{img} and E_{event} (blue) to process unpaired images and event frames. The applied loss constraints are visualized with red arrows. During inference, only the event encoder E_{event} and the task network T are required, both are computationally light-weight networks.

2.1. Event Generation based on Pseudo-Flow

Following the event generation model [4], see Eq. 2, translated events can be generated by using the image gradient $\nabla\tilde{I}_{(x,y)}$ and optical flow.

$$\begin{aligned}\Delta\tilde{I}_{(x,y)} &\approx -\langle\nabla\tilde{I}_{(x,y)}, \mathbf{v}_{(x,y)}\Delta t\rangle \\ &= -|\nabla\tilde{I}_{(x,y)}||\mathbf{v}_{(x,y)}|\Delta t\cos\alpha\end{aligned} \tag{2}$$

Instead of optical flow, we propose to directly predict pseudo-flow vectors $\hat{\mathbf{v}}_{(x,y)}$, which implicitly contain the unknown parameters Δt, $\cos\alpha$ and C. Thus, we do not need to compute these parameters explicitly. Our pseudo-flow is not equivalent to optical flow as the adversarial training only enforces realistic events by either adjusting the direction or the magnitude of $\hat{\mathbf{v}}_{(x,y)}$. The resulting pseudo-flow field adheres to the content extracted from an image $\mathbf{z}_{\text{img}}$ but with the general motion information of the event data, encoded in the *sensor-specific* feature ζ_{event}.

3. Experiments

We validate our approach for event classification on the Neuromorphic-Caltech101 (N-Caltech101) [11] dataset using the image-based Caltech101 as a labeled source dataset. For object detection, we train on the labeled image dataset Waymo Open Dataset [16] and evaluate on the Multi-Vehicle Stereo Event Camera Dataset (MVSEC) [18]. While several modules are used during training, crucially, during testing, we only use a fast ResNet-18 backbone. The event histogram [10] is used as event representation to facilitate the image-to-event translation.

Results The classification accuracies are reported in Tab. 1. Our approach outperforms the state-the-art method E2VID by 2.7% in terms of accuracy. Moreover, our inference network is a simple Resnet18, which is computationally much more lightweight than E2VID. Because of the increased size of the training dataset, our approach even outperforms the supervised methods. Moreover, the significantly lower performance of a simple cycle translation UDA framework shows that the feature space split is crucial for the task transfer between images and events. On the task of object detection, we compare against the event simulator ESIM [12] as well as E2VID. Additionally, EventGAN [19] is included as a paired baseline, i.e., it was trained with events and the corresponding frames. The object detection performances on MVSEC are reported in Table 2 as mean Average Precision (mAP) [1]. Compared to approaches trained on unpaired data, our approach achieves the highest performance, outperforming the next best method [13] by 26% in terms of mAP.

Method	Setting	Accuracy ↑
E2VID [14]	UDA	0.821
VID2E [5]	UDA	0.807
Simple Cycle	UDA	0.577
Ours	UDA	**0.848**
E2VID [14]	Supervised	0.866
VID2E [5]	Supervised	0.906
EST [6]	Supervised	0.817
HATS [15]	Supervised	0.642
Ours supervised	Supervised	0.839
EvDistill* [17]	UDA	0.902
Ours*	UDA	0.938

Table 1. Classification accuracies on the N-Caltech101 dataset. To stay consistent with the evaluation in [17], we report the performance achieved by our model trained on the whole Caltech101 dataset(*).

Method	Unpaired	mAP ↑
ESIM [12]	✔	0.02
E2VID [14]	✔	0.28
Ours	✔	**0.54**
EventGAN [19]	✗	0.30

Table 2. Mean average precision for the task of object detection on the MVSEC dataset.

References

[1] Mark Everingham, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 06 2010.

[2] Robert Fisher. Cvonline: Image databases. https://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm, 2021.

[3] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

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

[5] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to Events: Recycling video datasets for event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020.

[6] Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Int. Conf. Comput. Vis. (ICCV)*, 2019.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Conf. Neural Inf. Process. Syst. (NIPS)*, pages 2672–2680, 2014.

[8] Yuhuang Hu, Tobi Delbruck, and Shih-Chii Liu. Learning to exploit multiple vision modalitiesby using grafted networks. In *Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5967–5976, 2017.

[10] 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5419–5427, 2018.

[11] Garrick Orchard, Ajinkya Jayawant, Gregory K. Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Front. Neurosci.*, 9:437, 2015.

[12] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. In *Conf. on Robotics Learning (CoRL)*, 2018.

[13] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019.

[14] 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.*, 2019.

[15] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. HATS: Histograms of averaged time surfaces for robust event-based object classification. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1731–1740, 2018.

[16] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[17] 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021.

[18] Alex Zihao Zhu, Dinesh Thakur, Tolga Ozaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. *IEEE Robot. Autom. Lett.*, 3(3):2032–2039, July 2018.

[19] Alex Zihao Zhu, Ziyun Wang, Kaung Khant, and Kostas Daniilidis. Eventgan: Leveraging large scale image datasets for event cameras. *arXiv preprint arXiv:1912.01584*, 2019.