N-ROD: a Neuromorphic Dataset for Synthetic-to-Real Domain Adaptation

Marco Cannici[*,1] Chiara Plizzari[*,2] Mirco Planamente[*,2,3] Marco Ciccone[1]
Andrea Bottino[2] Barbara Caputo[2,3] Matteo Matteucci[1]

[1] Politecnico di Milano
name.surname@polimi.it

[2] Politecnico di Torino
name.surname@polito.it

[3] Istituto Italiano di Tecnologia
name.surname@iit.it

Abstract

Event cameras are novel neuromorphic sensors, which asynchronously capture pixel-level intensity changes in the form of "events". Event simulation from existing RGB datasets is commonly used to overcome the need of large amount of annotated data, which lacks due to the novelty of the event sensors. In this context, the possibility of using event simulation in synthetic scenarios, where data generation is not limited to pre-existing datasets, is to date still unexplored. In this work, we analyze the synth-to-real domain shift in event data, i.e., the gap arising between simulated events obtained from synthetic renderings and those captured with a real camera on real images. To this purpose, we extend to the event modality the popular RGB-D Object Dataset (ROD), which already comes with its synthetic version (SynROD). The resulting Neuromorphic ROD dataset (N-ROD) is the first to enable a synth-to-real analysis on event data, showing the effectiveness of Domain Adaptation techniques in reducing the synth-to-real shift. Moreover, through extensive experiments on multi-modal RGB-E data, we show that events can be effectively combined with conventional visual information, encouraging further research in this area. The N-ROD dataset is available at https://N-ROD-dataset.github.io/home/.

1. Introduction

Event cameras are neuromorphic bio-inspired vision devices that asynchronously stream events in correspondence of pixels subject to brightness changes. Despite the sensor's benefits (such as memory efficiency, time resolution, and dynamic range), the main challenge when training deep event-based architectures is the lack of annotated data. A common option to mitigate this issue is to use data simulation as an alternative to direct training on real data, as it provides a way to re-purpose existing RGB datasets in the event domain. However, the simulations do not always match the data distribution coming from a real sensor [11, 30]. As a result, training on simulated data could lead to sub-optimal performance. In this setting, the authors of [23] refer to this problem as the *sim-to-real* domain gap and propose to address it using unsupervised domain adaptation techniques.

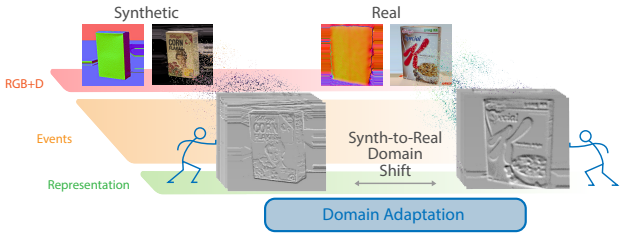

Figure 1. *How can we study the Synth-to-Real gap in event-based cameras?* We propose N-ROD, a new dataset explicitly designed for supporting research in domain adaptive event-based classification, in both single and multi-modal settings.

Moreover, simulating events from frames inherits all the complexities and costs associated with standard data collection and becomes inexpensive only if the RGB dataset is already available. Indeed, collecting data with precise annotation is a hard problem even with standard vision devices. In the literature, a common solution is to use synthetic data generation, as it provide free access to precise annotations. Nevertheless, differences between synthetic training data and real test one, commonly referred to as the *synth-to-real* domain shift, severely undermine the final model's performance on the actual data. Domain adaptation techniques revealed once more to be a powerful way to cope with this issue [4, 32, 27]. However, to date, due to the lack of suitable datasets for such an analysis, the real impact of the *synth-to-real* shift on events remains an open problem.

In this work, we propose to address this absense by extending the popular RGB-D Object Dataset (ROD) [16] for object recognition by its event counterpart. ROD comes with the RGB and depth modalities, both acquired through real sensors, and it was recently extended with synthetic samples [19]. We further extend ROD, and its synthetic version SynROD, by introducing event data to enable the

[*]The authors equally contributed to this work.

synth-to-real analysis for the event modality, resulting in a new neuromorphic dataset which we call N-ROD. Thanks to the multiple modalities already provided by the original ROD, the new N-ROD does not only enable the study of the domain gap between synthetic and real data, but it also unlocks the possibility of designing new ways of exploiting events together with conventional vision information, encouraging further research in this direction.

2. Related Works

Datasets. Early works in event-based vision focused in recognizing simple objects and shapes such as digits [22, 28] and poker pips [28]. A standard procedure to re-purpose existing RGB dataset is to record them using a real camera, which enabled the creation of more complex datasets [17, 22, 15]. Thanks to recent advances in event research, the availability of more realistic datasets recorded in real conditions [29] and more complex vision understanding tasks [3, 2] is also increased. Contrary to all previous datasets on classification, however, the one proposed in this work is the first to enable domain adaptive analysis.

Unsupervised Domain Adaptation (UDA). Unsupervised domain adaptation focuses on bridging the gap between a labeled source domain and an unlabeled target one by acting on reducing the difference between the feature spaces of the two. Multiple approaches are possible, from those *directly* acting on features, such as discrepancy-based methods [35, 26, 20, 21, 7] and those based on adversarial training [8, 31, 10, 25], to those that *indirectly* act on the feature spaces by means of pretext tasks [19, 34, 6, 5]. The research in the multi-modal field started from simple applications of existing single-modal DA methods [33, 18], and is now moving to more mature approaches which specifically exploit the multi-modal nature of the data [19].

UDA for events. Very new is UDA in event-based data. Its effectiveness in tackling the gap between real events and simulated ones was recently shown in [23], which deviates from previous approaches acting on simulation parameters [30] and data augmentation [11]. Moreover, the authors of [23] show that, when RGB is the primary modality, adding simulated events obtained from both synthetic source images and real target ones can be used as a tool in UDA to mitigate the intrinsic synth-to-real shift *on images*. Our work differs from [23] in that it shifts the focus to how real event data benefits from DA techniques, extending the synth-to-real shift *on events* thanks to the proposed N-ROD dataset. Under this setting, simulation is not performed on the target domain as real events are already available.

3. Dataset

We propose an extension of the popular RGB-D Object Dataset (ROD) [16] for object recognition. ROD contains

Synthetic Real

Figure 2. Synthetic (left) and real (right) samples from the N-ROD dataset. Depth images are colorized with surface normal encoding and event sequences are represented using voxelgrid [36].

41,877 samples of 300 daily objects grouped into 51 categories, captured by an RGB-D camera. ROD is coupled with SynROD [19], its recent synthetic extension designed to study the synth-to-real domain shift in multi-modal settings, i.e., RGB images and depth. SynROD contains photorealistic renderings generated under natural lighting conditions of 3D models of the same categories as ROD.

In this work, we extend both versions of the dataset by introducing real event recordings obtained from ROD samples, as well as simulated events extracted from SynROD's synthetic images. The resulting extended dataset is the first to enable a synth-to-real analysis on event data.

Recording Setup. We replicate the setting in [22] for converting RGB images to event-based recordings. A Prophesee's HVGA Gen3 (CD+EM) [9] Asynchronous Time Based Image Sensor (ATIS) configured with default bias settings and mounting a Computar M0814-MP2 8mm lens is placed on a pant-tilt and positioned at approximately 23 centimeters from a LCD monitor. We used a 2560×1440 76Hz IPS monitor with a 4ms minimum response time (Lenovo^{TM} ThinkVision® P27h-10), and set its brightness and contrast settings to their highest values as in [15]. The pan-tilt[1], analogous to the one used in [22], is composed of two Dynamixel MX-28 servo motors connected with each other, and an ArbotiX-M Robocontroller board controls it through serial communication.

Objects from the ROD dataset come into crops of variable size and aspect ratio. Samples are processed with padding, which replicates the border on the smallest side of the image to ensure squared samples despite the original resolution. We display still images from the original ROD dataset in a loop and record each sample while performing the same saccades motion pattern described in [22] (i.e., three saccades motions of 100ms each in a triangular pattern). A 300ms waiting time was added after transitioning to the next image to ensure the image was correctly updated on the monitor, and the event camera has settled after detecting the visual changes incurred by changing the image. A 256×256 region of interest was set on the event cam-

[1] https://trossenrobotics.com/widowx-MX-28-pan-tilt

era to restrict recorded events to a squared resolution, as in ROD RGB images. Grayscale images from exposure measurement (EM) events were used to fit the size of displayed images to the camera field of view before recording.

To simulate data on the source domain we follow the procedure used in [11, 23], making use of the ESIM [24] simulator to generate events. We replicate the same setting used to record real samples, mapping synthetic images on a plane and moving the virtual event camera with the usual saccadic motion.

4. Method

The proposed N-ROD dataset enables the analysis of the synth-to-real domain shift in event-based data. In this setting, the source comprises pairs of synthetic RGB images and their event version generated using a simulator [24]. The target domain instead includes both RGB images and events acquired with real sensors. Thus, the N-ROD setting is different from the one in [23], where event simulation is applied on both the source and target domains. Simultaneously, using simulation on one side and real event data on the other, N-ROD indirectly introduces the sim-to-real gap studied in [11, 30]. The result is a *double domain-shift*, which combines both the synth-to-real and the sim-to-real shifts. To cope with both these aspects, we use the UDA algorithms analyzed in the next section.

4.1. UDA Algorithms

In this section we give a brief overview of the UDA methods applied within our architecture.

GRL. Ganin *et al.* [10] introduced a domain adversarial method exploiting a *gradient reversal layer* which ensures that the feature distributions of the two domains become indistinguishable (domain-invariant) to the feature extractor.

MMD. Long *et al.* [20] proposed to minimize the Maximum Mean Discrepancy between source and target distributions, a metric that measures the discrepancy between them. By doing so, the final layers of the network are encouraged to produce domain-invariant features.

AFN. Xu *et al.* [35] pointed out that the main difficulty in classifying on target domain is due to target vectors having smaller feature norms if compared to that of the source domain. To tackle this issue, the authors proposed to iteratively increase the expectations of the L_2-norms of the deep embeddings of source and target domains.

Rotation. Xu *et al.* [34] proposed to add an auxiliary self-supervised task to the main loss, which consist in predicting the absolute rotation of images from both the source and target domains. Loghmani *et al.* [19] extended it to multi-modal images, asking the network to predict the *relative rotation* between two modalities.

Entropy. Grandvalet *et al.* [13] proposed to represent the uncertainty on the target domain and add a regularization term to the classification loss that helps soften the domain shift effects between source and target distributions.

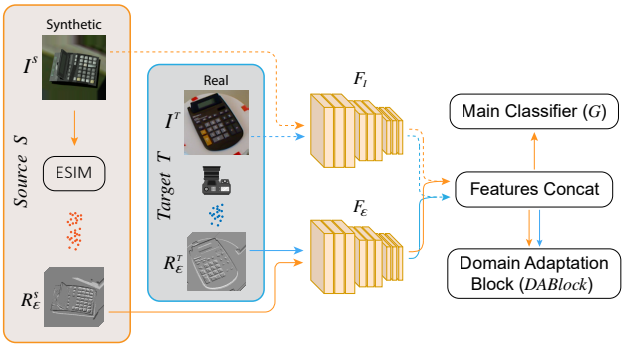

Figure 3. Our multi-modal DA architecture. Data coming from the source and target domains are processed separately during training. Source, labelled, data is used for supervised classification in G, while both target and source data are fed to the *DABlock*. Features are extracted from each modality using different extractors F_I and F_ϵ, shared across domains, and then concatenated before prediction. The dashed data path is finally removed, along with features concatenation, when just the event modality is used.

4.2. Network Architecture

In Figure 3 we outline the structure of the proposed network. We indicate with I^S images from the source domain and with I^T those from the target. Given the sets of input RGB images (I^S, I^T), events in the source domain are obtained from the ESIM simulator [24], and that in the target domain using an event-based camera. Event data is then split into $B = 9$ evenly spaced temporal bins from which independent voxel-grids are extracted as described in [36], resulting in a multi-channel representation $(\mathcal{R}^S_\mathcal{E}, \mathcal{R}^T_\mathcal{E})$. Both the RGB and event inputs are then fed to two separate ResNet feature extractors (respectively, F_I and $F_\mathcal{E}$), which are shared between source and target. The first convolutional layer of $F_\mathcal{E}$ is replaced with a new one matching the B input channels and randomly initialized. Both source and target features are fed to the DABlock that implements one of the techniques presented in Section 4.1, while the final classifier G is trained on the source domain features only.

Implementation Details. The two backbones F_ϵ and F_I are implemented with a ResNet18 [14] pretrained on ImageNet. Both voxel-grid representations and RGB images are augmented during training following [23]. We train all network configurations for 30 epochs using SGD as optimizer and weight decay 0.003.

5. Experimental Results

In this evaluation we focus specifically on how UDA techniques are effective in reducing the synth-to-real domain shift, by analysing their impact first on event data individually, and then in a multi-modal setting.

Single modality Benchmark. Table 1 reports top-1 accuracy results obtained in the synth-to-real scenario, where the source is Syn-N-ROD (i.e., events simulated on synthetic data) and the target is N-ROD (i.e., events captured with a neuromorphic camera from real RGB images). For each modality, the baseline results are referred to as "Source Only" and they are obtained by training on source and testing on target samples without applying any UDA approach. We remark that, in this setting the RGB modality is favored by robust pre-trained layers. Results show that, among all modalities, the event one is that receiving the highest benefits from UDA (20.6% compared to Source-Only, while RGB and depth have smaller improvements of, respectively, 9.9% and 14.4%). The event modality focuses on geometric components and object shapes, contrary to the RGB which is biased towards texture. This intrinsic peculiarities make UDA techniques more effective on events, as shape information is per se more robust [12] in the transition from the synthetic to real domain and thus easier to be aligned, as opposed to the information encoded in RGB.

Multi-modal Benchmark. It is well known that the complementarity of different input modalities, such as RGB and depth, can be exploited to improve adaptation performance in cross-domain scenarios [19]. Since a multi-modal RGB-E analysis is still unexplored in the literature, we propose a first approach to the problem relying on the one commonly used for RGB-D data [1] (see Section 4.2). The results in Table 1 show that the validity of the DA approach is confirmed in the multi-modal setting, as all methods consistently improve over the Source Only baseline, as in the single modal setting. *Rotation* [34] provides more interesting results to be discussed. Indeed, when absolute rotation is applied on each modality individually, this method is the one achieving the lower performance gain if compared to all the others. Instead, when extended to the RGB-E context by applying the *Relative Rotation* [19] between the two modalities, it interestingly reveals to be the UDA method performing the best. This brings to light the importance of leveraging over the complementarity of multiple modalities even in the event field. This result emphasizes the need to further push research towards networks specifically designed to make the two modalities efficiently cooperate.

Synth-to-Real *vs* Sim-to-Real. Using simulation on one side and real event data on the other indirectly introduces the sim-to-real gap studied in [11, 30]. In order to understand how much this second domain shift affects performance, we compare our results with the ones which would been obtained by using simulation even to extract events from real (target) image. We can consider this experiment as a relaxed setting where the sim-to-real gap is not present. To this purpose, Table 2 compares our single-modal results obtained by generating target events with a neuromorphic camera (Source: Sim, Target: Real) with the ones obtained through simulation (Source: Sim, Target: Sim).

Table 1. Top-1 accuracy (%) of UDA methods on synth-to-real shift (Syn-N-ROD → N-ROD). **Bold:** highest mean result, underline: highest single- and multi-modal results. ▲ indicates the improvement of the avg of UDA methods over the baseline Source Only.

SYNTH-N-ROD ⟹ N-ROD					
	Single-modal			Multi-modal	
Method	RGB	Depth	**Event**	RGB+D	**RGB+E**
Source Only	52.13	7.56	21.78	47.70	50.78
GRL [10]	57.12	26.11	33.09	59.51	57.15
MMD [20]	63.68	29.34	42.05	62.57	61.78
Rot [34][19]	63.21	6.70	31.26	66.68	68.54
AFN [35]	64.63	30.72	55.12	62.40	64.04
Entropy [13]	61.53	16.79	50.14	63.12	64.08
Avg	62.03 ▲+9.9	21.93 ▲+14.4	42.33 ▲+20.6	62.86 ▲+15.2	**63.12** ▲+12.3

Table 2. Top-1 accuracy (%) on events, in two different scenarios: *sim-to-real* and *sim-to-sim*. In **bold** the highest mean result.

SYNTH-N-ROD ⟹ N-ROD							
Source	Target	Source Only	GRL	MMD	Rot	AFN	Entropy	Avg
Sim	Real	21.78	33.09	42.05	31.26	55.12	50.14	42.33
Sim	Sim	40.47	44.52	48.29	42.98	53.50	49.29	**47.68**

By considering the results on Source Only, we notice that, without any kind of adaptation, performance decreases by up to 20%. This quantifies the sim-to-real gap and highlights the need of a method to reduce the gap between simulated and real data. The proposed approach once again reveals the effectiveness of using UDA techniques in the event context, which consistently improve performance, reducing this gap to be only 5%.

6. Conclusions

This paper presents N-ROD, a new neuromorphic dataset explicitly designed for supporting research in domain adaptive event-based object classification. This dataset opens new research opportunities for studying both the synth-to-real and sim-to-real domain gaps. Moreover, we believe that this work will become a valuable starting point for the community also for further research towards new ways of integrating event data with other modalities.

Acknowledgements. The work was partially supported by the ERC project N. 637076 RoboExNovo. We also acknowledge that the research activity herein was carried out using the IIT HPC infrastructure.

References

[1] Andreas Aakerberg, Kamal Nasrollahi, Christoffer Bøgelund Rasmussen, and Thomas B Moeslund. Depth value preprocessing for accurate transfer learning based rgb-d object recognition. In *International Joint Conference on Computational Intelligence*, pages 121–128. SCITEPRESS Digital Library, 2017. 4

[2] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael Debole, Steve Esser, Tobi Delbruck, Myron Flickner, and Dharmendra Modha. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[3] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based object classification for neuromorphic vision sensing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[4] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 1

[5] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *arXiv preprint arXiv:1608.06019*, 2016. 2

[6] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 2

[7] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulò. Just dial: domain alignment layers for unsupervised domain adaptation. In *ICIAP*, 2017. 2

[8] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9944–9953, 2019. 2

[9] G. Gallego, T. Delbruck, G. M. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 2

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015. PMLR. 2, 3, 4

[11] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3586–3595, 2020. 1, 2, 3, 4

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 4

[13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, page 529–536, Cambridge, MA, USA, 2004. MIT Press. 3, 4

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[15] 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. 2

[16] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *2011 IEEE international conference on robotics and automation*, pages 1817–1824. IEEE, 2011. 1, 2

[17] 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

[18] Xiao Li, Min Fang, Ju-Jie Zhang, and Jinqiao Wu. Domain adaptation from rgb-d to rgb images. *Signal Process.*, 131(C):27–35, Feb. 2017. 2

[19] Mohammad Reza Loghmani, Luca Robbiano, Mirco Planamente, Kiru Park, Barbara Caputo, and Markus Vincze. Unsupervised domain adaptation through inter-modal rotation for rgb-d object recognition. *IEEE Robotics and Automation Letters*, 5(4):6631–6638, 2020. 1, 2, 3, 4

[20] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 2, 3, 4

[21] Massimiliano Mancini, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Adagraph: Unifying predictive and continuous domain adaptation through graphs. In *CVPR*, 2019. 2

[22] Garrick Orchard, Ajinkya Jayawant, Gregory K. Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9(NOV), Jan. 2015. 2

[23] Mirco Planamente, Chiara Plizzari, Marco Cannici, Marco Ciccone, Francesco Strada, Andrea Bottino, Matteo Matteucci, and Barbara Caputo. Da4event: towards bridging the sim-to-real gap for event cameras using domain adaptation. *arXiv*, 2021. 1, 2, 3

[24] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on Robot Learning*, pages 969–982. PMLR, 2018. 3

[25] Paolo Russo, Fabio Maria Carlucci, Tatiana Tommasi, and Barbara Caputo. From source to target and back: symmetric bi-directional adaptive gan. In *CVPR*, 2018. 2

[26] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Con-*

ference on Computer Vision and Pattern Recognition, pages 3723–3732, 2018. 2

[27] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018. 1

[28] Teresa Serrano-Gotarredona and Bernabé Linares-Barranco. Poker-dvs and mnist-dvs. their history, how they were made, and other details. *Frontiers in neuroscience*, 9:481, 2015. 2

[29] 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

[30] 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 *Proc. Eur. Conf. Comput. Vis.* Springer, 2020. 1, 2, 3, 4

[31] Hui Tang and Kui Jia. Discriminative adversarial domain adaptation. In *AAAI*, pages 5940–5947, 2020. 2

[32] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7364–7373, 2019. 1

[33] Jing Wang and Kuangen Zhang. Unsupervised domain adaptation learning algorithm for rgb-d staircase recognition. *ArXiv*, abs/1903.01212, 2019. 2

[34] Jiaolong Xu, Liang Xiao, and Antonio M. Lopez. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7:156694–156706, 2019. 2, 3, 4

[35] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 3, 4

[36] 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. 2, 3