E^2(GO)MOTION: Motion Augmented Event Stream for Egocentric Action Recognition

Chiara Plizzari[*,1] Mirco Planamente[*,1,2] Gabriele Goletto[1] Marco Cannici[3] Emanuele Gusso[1]
Matteo Matteucci[3] Barbara Caputo[1,2]

[1] Politecnico di Torino [2] CINI Consortium [3] Politecnico di Milano

name.surname@polito.it name.surname@polimi.it

Abstract

Event cameras are novel bio-inspired sensors, which asynchronously capture pixel-level intensity changes in the form of "events". Due to their sensing mechanism, event cameras have little to no motion blur, a very high temporal resolution and require significantly less power and memory than traditional frame-based cameras. These characteristics make them a perfect fit to several real-world applications such as egocentric action recognition on wearable devices, where fast camera motion and limited power challenge traditional vision sensors. However, the ever-growing field of event-based vision has, to date, overlooked the potential of event cameras in such applications. In this paper, we show that event data is a very valuable modality for egocentric action recognition. To do so, we introduce N-EPIC-Kitchens, the first event-based camera extension of the large-scale EPIC-Kitchens dataset. In this context, we propose two strategies: (i) directly processing event-camera data with traditional video-processing architectures (E^2(GO)) and (ii) using event-data to distill optical flow information (E^2(GO)MO). On our proposed benchmark, we show that event data provides a comparable performance to RGB and optical flow, yet without any additional flow computation at deploy time, and an improved performance of up to 4% with respect to RGB only information. The N-EPIC-Kitchens dataset is available at https://github.com/EgocentricVision/N-EPIC-Kitchens.

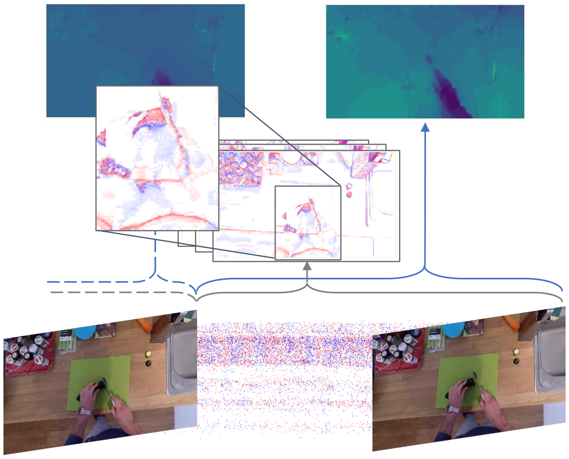

Figure 1. **N-EPIC-Kitchens**: the first event-based dataset for egocentric action recognition. From RGB images, we generate a stream of events (bottom). Positive polarity is represented by red events, whereas blue events represent negative polarity. Events focus on motion, similarly to optical flow (top). With their low latency, high temporal resolution, and low-power consumption, event data are a perfect fit for egocentric action recognition.

1. Introduction

Egocentric vision has introduced a variety of new challenges to the computer vision community, such as human-object interaction [18,65], action anticipation [1,30,39,64], action recognition [52], and video summarization [23,57,58]. With the advent of novel large-scale datasets [14,15], new tasks are being proposed, such as wearer's pose estimation [105] and egocentric videos anonymization [95]. This trend will grow in the next years thanks to the very recent release of Ego4D [41], a massive-scale egocentric video dataset offering more than 3,000 hours of daily-life activity videos accompanied by audio, 3D meshes of the environment, eye gaze, stereo, and multi-view videos.

Among all, RGB sensors provide by far the richest source of visual information. However, the performance of RGB-based deep models drastically decrease when the training and test data do not share the same distribution [20]. This issue, known as environmental bias [53,72,78,85,89], originates from RGB-based networks' tendency to rely on the environment in which activities are recorded, affecting their ability to recognize actions when they are performed in unfamiliar (unseen) surroundings. This is mainly caused by appearance-based networks' tendency to primarily focus on background cues and objects texture, which are typically uncorrelated with the action being performed and thus largely varying in different environments. As a result, appearance-free modalities, such as motion, have become the favored choice in current egocentric vision systems, as

[*] The authors equally contributed to this work.

testified by the results of recent EPIC-Kitchens challenges [16, 17, 19]. However, the optical flow used in this setting is computed from RGB frames by solving expensive optimization problems (TV-L1 algorithm [108]), introducing significant test-time computations [12].

Event-based cameras, on the other hand, have been shown to be particularly suitable for online settings [24,31]. Their high pixel bandwidth results in reduced motion blur, and the extremely low latency and low power consumption make these novel sensors particularly good in egocentric scenarios, where fast motion often impacts RGB-based systems negatively. Moreover, as they only convey differential information, event sequences reveal more information about the dynamic of the scene than its appearance, making them a valid alternative to RGB frames when learning to focus on motion. Still, despite these advantages, no prior research has looked at how to exploit their sensitivity to motion in egocentric vision, where these devices remain unused.

As a first step in this direction, we propose N-EPIC-Kitchens, a novel dataset that enables, for the first time, the use of event data in this context. It consists in the extension of the large-scale EPIC-Kitchens dataset [14] under the setup proposed in [72]. The latter is particularly appealing for both the availability of multiple environments (kitchens) and multiple modalities, i.e., RGB, optical flow, and audio. These characteristics allow for the analysis of the aforementioned environmental bias as well as the comparison of event data to well-established modalities. On the proposed N-EPIC-Kitchens, we introduce two approaches to exploit the intrinsic motion characteristics of event data in this context. The first, which we call E^2(GO), consists in extending traditional 2D and 3D action recognition architectures with layer variations aimed at exploiting the motion-rich features of event data. The second, E^2(GO)MO, extends motion reasoning by distilling motion information from optical flow to event data. This is accomplished following a teacher-student approach that allows taking full advantage of expensive offline TV-L1 flow during training only, while avoiding its computation at test time. We summarize our contributions as follows:

- We release N-EPIC-Kitchens, the first event-based egocentric action recognition dataset, which unlocks the possibility to explore event data in this context;
- We benchmark N-EPIC-Kitchens on popular action recognition architectures, showing performance of both event data alone and combined with RGB and optical flow modalities. Moreover, we demonstrate the robustness of event data to environment changes;
- We propose E^2(GO) and E^2(GO)MO, two event-based approaches tailored at emphasizing motion information captured by event data in egocentric action recognition;
- We show that event data can outperform RGB in challenging unseen environments and are competitive with them in known environments, suggesting that using event data is a viable option and more research should be performed in this direction.

2. Related Works

Event-based Vision. Taking advantage of the event-based cameras' inherent ability to perceive changes [24,31], researchers have started creating new solutions to tackle traditional computer vision problems exploiting this new way of sensing the world, including optical flow prediction [37,113], motion segmentation [73,112], depth estimation [35, 44], and many others. While traditional cameras are capable of providing very rich visual information at the tradeoff of slow and often redundant updates, event-based cameras are asynchronous and spatially sparse, and capable of microseconds temporal resolution. Event-based systems range from designs that focus on exploiting and maintaining event-camera sparsity during computation [4,86,107], to algorithms that combine events with standard cameras [7,35, 46,79,99], exploiting the complementarity of the two. With the goal of achieving minimum-delay computing, research has also focused on asynchronous designs, either by modifying regular CNNs [5, 69] or by utilizing specific hardware solutions [2, 21, 29], often leveraging on bio-inspired computing frameworks [68]. Despite event-based cameras have already been applied to action and gesture recognition tasks [10, 48, 67], previous works have not taken advantage of their complementarity with other visual modalities yet in these contexts, and used these cameras mainly in controlled environments where both the camera and the background are static [3, 70]. In this paper, instead, we tackle egocentric action recognition with events for the first time and combine them with other modalities.

Action Recognition. The success of 2D CNNs in the context of image recognition [43, 49] inspired the first video understanding architectures. Traditional 2D CNNs are often used to process frames individually, eventually fusing optical flow information [103], while late fusion mechanisms ranging from average pooling [102], multilayer perceptrons [111], recurrent aggregation [26, 61], and attention [40, 92] are employed to model temporal relations for action understanding. The use of 3D convolutions has also been proposed as an alternative [8, 96]. However, despite their ability to learn spatial and temporal relations simultaneously, they often introduce more parameters, requiring pre-training on large-scale video datasets [8]. To reduce the model's complexity, other approaches focus on finding more efficient architectures [28, 82, 94, 97, 98, 106]. As an example, a parameter-free channel-wise temporal shift operator has been introduced in the Temporal Shift Module (TSM) network [62], resulting in a 2D CNN capable of

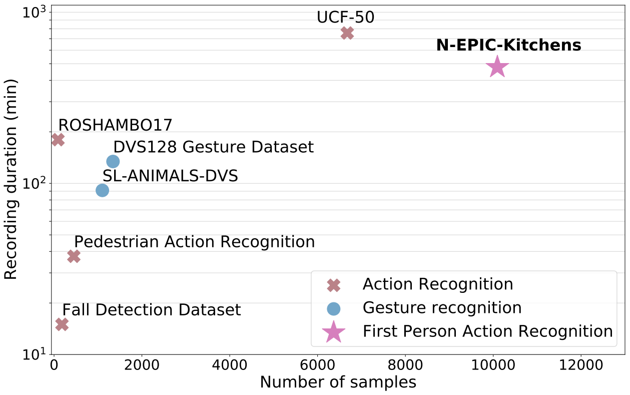

Figure 2. N-EPIC-Kitchens *vs* existing event-based action classification datasets in the literature [3, 47, 66, 70, 100].

encoding temporal information. Although all these architectures aim at implicitly modeling motion, most of them still mix video frames with the externally estimated optical flow. While this improves the overall performance, it also requires pre-computing the flow, making these approaches impracticable in online settings. In addition, two-stream approaches come at the cost of increased model complexity and number of parameters. To overcome this issue, a line of research proposes approaches that integrate the RGB and optical flow modalities in lighter architectures [56, 101, 110]. Finally, authors of [12, 77] proposed to distill optical flow information to the RGB stream at training time, while avoiding flow computation at test time.

First Person Action Recognition. The complex nature of egocentric videos raises a variety of challenges, such as ego-motion [60], partially visible or occluded objects, and environmental bias [53, 72, 78, 85, 89], which limit the performance of traditional, third-person, approaches when used in first person action recognition (FPAR) [14, 15]. The community's interest has quickly grown [16, 17, 19, 84] in recent years, thanks to the possibilities that these data open for the evaluation and understanding of human behavior, leading to the design of novel architectures [30, 51, 52, 92, 104]. While the use of optical flow has been the de-facto procedure [14–17, 19, 41] in FPAR, the interest has recently shifted towards more lightweight alternatives, such as gaze [27, 59, 71], audio [9, 52, 78], depth [32], skeleton [32], and inertial measurements [41], to enable motion modeling in online settings. These, when combined with traditional modalities, produce encouraging results, but not enough to make them viable alternatives. With this work, we show that the intrinsic motion information encoded by event data makes this modality potentially more suitable than RGB.

3. N-EPIC-Kitchens

Thanks to their focus on capturing only variations in the scene, event-based cameras are particularly efficient in egocentric scenarios, as they drastically reduce the amount of data to be processed and acquired, avoiding motion blur artifacts and providing fine-grained temporal information. However, so far only a limited amount of datasets have been made freely accessible [22, 36, 47, 75]. Despite the field is actively working towards increasing their availability, as testified by the recent release of event-based versions of ImageNet [54, 63], relatively few datasets for human activity recognition are currently available. As reported in Figure 2, most of them focus on action or gesture recognition [3, 47, 48, 70] in controlled settings, where both the camera and the background are static, and none considers egocentric action recognition, preventing event-based cameras use in this scenario. To demonstrate the advantages of event-based cameras in egocentric online settings, as well as their complementarity and equivalence to other modalities, we extend the EPIC-Kitchens (EK) [14] dataset, a large collection of egocentric action videos featuring multiple modalities and different environments. Following the setting of [72], we selected the three largest kitchens from EPIC-Kitchens in number of training action instances, which we refer to as D1, D2 and D3, analysing the performance for the 8 largest action classes, i.e., 'put', 'take', 'open', 'close', 'wash', 'cut', 'mix' and 'pour'.

In the following, we first introduce the operating principles of DVS cameras. Then, we outline the approach used to generate N-EPIC-Kitchens and emphasize its benefits.

3.1. Event-Based Vision Data

Pixels of DVS cameras are independent and respond to changes in the continuous log brightness signal $L(\mathbf{u}, t)$, differently from a standard RGB camera. An event is a tuple $e_k = (x_k, y_k, t_k, p_k)$ specifying the time t_k, the location (x_k, y_k) and the polarity $p_k \in \{-1, 1\}$ of the bright change (brightness decrease or decrease). An event is triggered when the magnitude of the log brightness at pixel $u = (x_k, y_k)^T$ and time t_k has changed by more than a threshold C since the last event at the same pixel, as described in the following equation:

$$\Delta L(\mathbf{u}, t_k) = L(\mathbf{u}, t_k) - L(\mathbf{u}, t_k - \Delta t_k) \geqslant p_k C. \quad (1)$$

Therefore, the output of an event camera is a continuous stream of events described as a sequence $\mathcal{E} = \{(x_k, y_k, t_k, p_k) | t_k \in \tau\}$, being τ the time interval.

N-EPIC-Kitchens generation. We leverage ESIM [83], a recent event camera simulator, to enhance the EPIC-Kitchens dataset with the event modality. Since videos in EPIC-Kitchens are limited to 60 frames per second, far lower than the microseconds temporal resolution of an event camera, we first upsample them to a higher fps. To this end, we used Super SloMo [50] for its unique ability to generate frames at any temporal precision, following the adap-

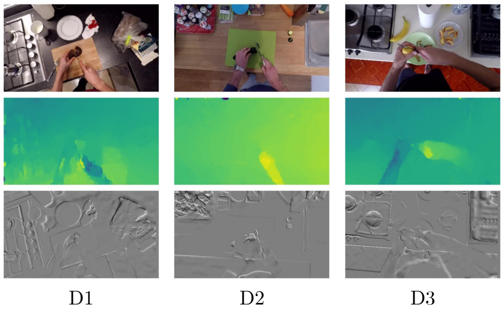

Figure 3. RGB (top), optical flow (middle) and Voxel Grid representation (bottom) from the same action ("cut") on the three different kitchens (D1, D2, D3).

tive sampling procedure proposed in Vid2E [33] to extract event streams. Finally, we use Voxel Grid [113], a frame-like event encoding technique, to convert sparse and asynchronous events to a tensor representation and enable learning with typical convolutional neural network architectures.

4. Challenges of Evaluating Event Data

The fundamental problem in assessing event data in first-person action recognition comes from the fact that, unlike other modalities, its use in egocentric vision is completely novel. To set a benchmark in this setting, we evaluate four different aspects of event-based modeling. We start by considering the importance of performance on both seen and unseen test sets, where *seen* indicates performance on the same kitchen on which training is performed, and *unseen* the performance obtained on a different one. We propose to evaluate them altogether in our experiments. While the first provides a good indication of the modality's upper bound performance, the second evaluates the ability of the model to encode domain invariant features and, as a result, the viability of using it in real-world scenarios. Then, as the performance of different modalities may greatly vary depending on the architecture used for processing [80], we benchmark events using three of the most accredited architectures in FPAR, namely TSM [62], TSN [103] and I3D [8]. We leverage a well-established procedure for converting event streams into a frame-like representation that has been shown to efficiently integrate with off-the-shelf CNNs [79,91], and finally propose to encourage modeling of motion features by employing attention at channel level.

Event Representation. Since event cameras produce sparse encodings of the scene, they must be converted into intermediate representations before processing. Several representations have been proposed, ranging from bio-inspired [5,11,68] to more practical ones. Frame-like representations are by far the most widespread methods as they can be directly used together with off-the-shelf networks. Among available ones [5,6,25,34,48,55,87,113] we chose Voxel Grid [113] as it proved to be superior in cross-domain settings [79,91]. This representation computes a B-channel image by discretizing time in B separate intervals:

$$\mathbf{x}^E(x,y,b) = \sum_{k=1}^{N} p_k k_b(b - t_k^*), \tag{2}$$

where b are the channels, t_k^* are the timestamps scaled into $[0, B-1]$, p_k is the polarity and $k_b(a) = max(0, 1-|a|)$.

Backbone Architectures. To assess how event data behaves on different network designs, we examine two popular 2D-CNN approaches, TSM [62] and TSN [103] as well as one 3D-CNN, I3D [8]. The first two rely on a 2D-CNN backbone, but while TSN [103] can only leverage late fusion for temporal modeling, TSM [62] exploits *shift modules* to exchange channel information across adjacent frames. In contrast, I3D [8] is a pure 3D-CNN model, which *inflates* filters and pooling kernels into the temporal dimension. In the literature, there is currently no clear winner, as some modalities may react better with one technique than the other indiscriminately.

The Importance of Motion. Environmental biases are typically managed in egocentric vision systems by employing complementary, often appearance-free, modalities. Optical flow is generally the one performing the best in action recognition tasks [14,15,103], as (i) it helps focusing on the moving content, i.e., the action being performed, while (ii) preserving the edges of moving objects and (iii) ignoring background information. In this paper, we argue that event cameras' sensitivity to moving edges and ability to disregard static information only partially capture the three key features of optical flow listed above. In reality, as a result of the camera movement, these sensors still catch events in the background. This encourages us to learn from flow in order to improve our ability to filter out less discriminative data.

5. Learning from Motion

While a traditional RGB frame encodes static information only, frame-based representations used for event data also carry motion information on the channel dimension (see Section 4). Indeed, each temporal channel encodes the motion that occurs in the blind-time between a pair of standard frames of the video recording. We propose two different approaches to make standard CNNs able to exploit this information. The first, which we name E^2(GO), explicitly models temporal relationships by introducing channel operations that promote motion reasoning. The second, instead, uses a student-teacher strategy that we call E^2(GO)MO to encourage the network to extract motion features during training by utilizing a pre-trained optical flow based network. We detail the two approaches in the following.

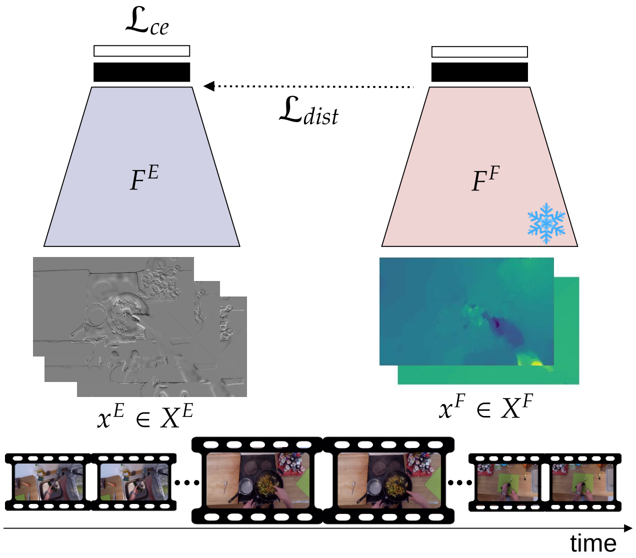

Figure 4. Illustration of the proposed E^2(GO)MO. The input $\mathbf{x}^E$ and $\mathbf{x}^F$ from the event and flow modality are passed to the feature extractors F^E and F^F respectively. Information from the pre-trained teacher stream (frozen) F^F is distilled to the student stream F^E. The latter is trained with standard cross-entropy loss.

5.1. E^2(GO): Event Motion

In order to enable standard CNNs to capture motion information from event data, we propose two simple but effective architectural variations, which improve the capability of extracting temporal inter-channel relations in 2D and 3D CNNs. We refer to them as E^2(GO)-2D and E^2(GO)-3D, respectively.

E^2(GO)-2D. A common practice in the literature is to extract temporal correlations at video level by modeling dependencies between different frames [52, 62]. A peculiarity of event representation is that the channel sequence encodes continuous motion, thus describing micro-movements in the scene. This observation motivates us to extend the practice of modeling temporal relations to also learn short-range correlations between event channels.

We propose to do this by exploiting *Squeeze And Excitation* modules [45] to enhance attention correlations between channels in 2D CNNs. Given an event volume $\mathbf{x}^E \in \mathbb{R}^{T\times H\times W\times F}$ as input, where T is the temporal dimension, $H \times W$ is the feature map resolution and F indicates the number of channels, we refer as $\mathbf{f}_i^E \in \mathbb{R}^{T\times H_i\times W_i\times C_i}$ to the features extracted from the i-th layer of the network. As a first step, we "squeeze" the spatial information content of $\mathbf{f}_i^E$ into a channel descriptor by performing feature aggregation along the spatial dimensions. It follows an "excitation" operator, which takes in input $\mathbf{z}_{sq}^E$ to produce an activation vector $\mathbf{s}$ to be used to scale $\mathbf{x}^E$. The scaling vector $\mathbf{s}$ is obtained from $\mathbf{z}_{sq}^E$ through two fully-connected layers with a bottleneck that down sizes C to C/r. Finally, $\mathbf{s}$ is used to re-weight $\mathbf{x}^E$, resulting in a new feature vector $\tilde{\mathbf{x}}^E$ to enhance discriminative motion features and discard the less informative ones. As a result, $\tilde{\mathbf{x}}^E$ encodes the relation dynamics between different temporal channels, effectively modeling the dependencies between them as a result of a self-attention function on channel dimension.

E^2(GO)-3D. Similarly, we propose to exploit 3D-CNNs' ability to process temporal information through a 3D kernel. Starting from the same input $\mathbf{x}^E \in \mathbb{R}^{T\times H\times W\times F}$, traditional 3D CNNs apply a 3D convolution on the (T, H, W, F) dimensions, resulting in an output of shape (T', H', W', C). We re-purpose the 3D convolution operator in this context to operate on $\mathbf{x}^E \in \mathbb{R}^{(F\cdot T)\times H\times W\times 1}$ by moving the channel dimensions on the temporal axis. The convolution directly models the micro-movements contained across the temporal channels of the event representation, which would otherwise be ignored when processed on the channel dimension.

5.2. E^2(GO)MO: Learning from Flow

Our goal is to train a network using both event and optical flow data, avoiding the need to estimate the latter during testing. Given a multi-modal input $X = (X^E, X^F)$, where X^E denotes the event modality and X^F denotes the flow one, we indicate with F^E and F^F their respective feature extractors, and the resulting features with $\mathbf{f}^E = F^E(\mathbf{x}^E)$ and $\mathbf{f}^F = F^F(\mathbf{x}^F)$. As a first step, we train the flow extractor F^F using a cross-entropy loss between the true action labels $\hat{y}$ and the labels y^F predicted by a fully connected layer on top of F^F. Then, we first freeze the flow stream F^F, and then train the event stream F^E by combining the standard cross-entropy loss with a *distillation loss* defined as the L_2 between features $\mathbf{f}^E$ and $\mathbf{f}^F$:

$$\mathcal{L}_{dist} = \alpha||\mathbf{f}^E - \mathbf{f}^F||^2. \tag{3}$$

where α is a scaling hyperparameter. Such loss encourages features of the event stream to match those of the flow one, forcing F^E to mimic the behavior of F^F, and thus enabling the two to produce similar activations. Notice that we use optical flow data only during training and remove the teacher branch during inference, thus exploiting the advantages of this modality but effectively avoiding its computational complexity in prediction.

6. Experiments

In this section, we first introduce the experimental setup used (Section 6.1), then we benchmark event data and validate the proposed E^2(GO) and E^2(GO)MO. We conclude the section with a discussion and limitation paragraph.

6.1. Experimental Setup

Input. Experiments with I3D [8] are conducted by sampling one random clip from the video during training and 5 equidistant clips spanning across all the video during test, as in [72]. The number of frames composing each clip is 16 for RGB and optical flow, and 10 for events. For TSN [103] and TSM [62] architectures, uniform sampling is used, consisting of 5 frames uniformly sampled along the video. During testing, 5 clips per video are adopted, following [62]. The Voxel Grid representations are clipped between -0.5 and 0.5, and all data modalities are rescaled and normalized in accordance with the pretrained network associated with the architecture adopted. For all modalities, we used standard data augmentation following [102].

Implementation and Training Details. With regard to I3D, the original implementation from [8] has been chosen, while TSN and TSM models have been built using respectively a BN-Inception [49] and a ResNet-50 [43] backbone. In the multi-modal experiments, a classic late fusion strategy is used, in which prediction scores from different modalities are summed and the error is backpropagated to all modalities. All models are implemented in PyTorch [74]. SGD with momentum [81] with a starting learning rate η of 0.01, a weight decay of 10^{-7} and a momentum μ of 0.9 is used as optimizer. We trained the networks for a total of 5000 iterations with a learning rate decay to $1e{-}3$ at step 3000. All the experiments are performed with a batch size of 128 on 4 NVIDIA Tesla V100 16Gb GPUs. For the distillation loss, we found the best hyperparameter $\alpha = 100$ (see Supplementary for details). As far as the evaluation protocol used, for *seen* results we train on kitchen D_i and test on the same ($D_i \rightarrow D_i$), $i \in \{1, 2, 3\}$. We evaluate performance on *unseen* test by training on D_i and testing on D_j, with $i \neq j$ and $i, j \in \{1, 2, 3\}$ ($D_i \rightarrow D_j$).

6.2. Results

Event Analysis. In Table 1 we show the performance of events on the three selected action recognition architectures (see Section 4). We observed that extracting 3-channels Voxel Grid is the optimal choice and we used it in all the remaining experiments (more details in Supplementary). Considering the performance on both seen and unseen test sets, the TSM model is the one performing the best, while I3D performs slightly worse. One explanation is that it only processes a small portion of the video at a time, catching only local features when trained at the clip level. TSM, on the other hand, can capture global features because it works with frames that cover the full video. The poor performance of TSN is to be expected, given that its frame aggregation prevents any temporal correlation from being modeled. Thus, unless otherwise stated, we perform video-level analysis and evaluate the proposed approaches on TSM and I3D backbones in all of the following experiments.

Model	Voxel ch.	Testing	Seen (%)	Unseen (%)
I3D	3	Clip	53.75	35.90
		Video	**55.54**	**37.52**
TSN	3	Clip	58.81	34.65
		Video	**59.82**	**35.24**
TSM	3	Clip	64.38	37.75
		Video	**65.93**	**38.23**

Table 1. *Mean* accuracy (%) over all $D_i \rightarrow D_j$ combinations on I3D, TSN and TSM on both seen and unseen test sets.

Event *vs* RGB. In Table 2 we compare events against the RGB modality. Results show that events surpass RGBs by up to 3% on unseen test sets. Indeed, it has been shown in the literature that appearance-based CNNs are biased toward texture, which causes them to underperform across-domain, but their robustness improves when shape-bias is increased [38]. We believe this is the primary reason why event representations, focusing more on geometric and temporal information rather than texture variations, are more invariant to domain changes. The same considerations also apply to seen tests, where RGB-based networks overfit by leveraging domain-specific features. We remark that until now the event modality was still lagging behind RGB images in purely visual tasks, as reported by the recent release of N-ImageNet benchmark [54], where the best performing event architecture scores 48.94%, considerably below RGB's $> 90\%$ accuracy [13, 42, 76, 109]. In this study, instead, we show that events can outperform RGBs in challenging unseen scenarios and compete under seen ones, emphasizing their importance in egocentric vision.

E^2(GO). In Table 2 we show the performance of E^2(GO)-2D and E^2(GO)-3D. Those are beneficial especially on unseen test sets, as they aim to enhance temporal correlations, thus allowing the network to emphasise motion features that are informative while suppressing those that are not correlated with the action. E^2(GO)-3D achieves an improvement by up to 2% on seen test set, while E^2(GO)-2D achieves results on-par with the baseline TSM. This could be explained by the fact that 2D CNNs, being based on frame-based techniques, rely heavily on visual signals. Indeed, while those are harmful when changing environments, they can be helpful on seen ones. I3D, on the other hand, is naturally more responsive to temporal correlations. Extending its temporal reasoning to micro-movements facilitates it in extracting discriminative features for the action, reflecting in an higher accuracy even when testing on the same environment.

Multi-Modal Analysis. Table 3 illustrates the behavior of the event modality when combined with RGB and optical

Modality	Model	D1	D2	D3	D1→ D2	D1→D3	D2→D1	D2→D3	D3→D1	D3→D2	Seen (%)	Unseen (%)
RGB	I3D	53.67	61.12	60.70	34.50	35.70	34.94	36.46	33.93	38.37	**58.49**	35.65
Event	I3D	50.32	58.33	57.99	37.27	39.12	32.98	36.52	35.68	43.56	55.54	37.52
Event	E^2(GO)-3D	50.52	62.99	60.11	38.07	38.71	35.02	38.49	36.73	45.53	57.87	**38.76**
RGB	TSM	61.61	77.08	75.75	37.39	32.49	34.28	38.99	34.43	38.25	**71.48**	35.97
Event	TSM	56.86	72.43	68.49	28.73	34.00	37.09	42.30	42.27	45.02	65.93	38.23
Event	E^2(GO)-2D	56.58	70.03	69.60	34.98	35.16	38.21	47.80	41.71	44.13	65.40	**40.33**

Table 2. Accuracy (%) of event w.r.t. RGB on both I3D and TSM. Results are shown on all shifts, i.e., $D_i \rightarrow D_j$ indicates we trained on D_i and tested on D_j, and D_i means we trained and test on the same. E^2(GO)-3D and E^2(GO)-2D improvements are shown w.r.t. to their respective baselines, where no architectural variations are performed. In **bold** the best results on both seen and unseen for each backbone.

Model	Streams	Pretrain	Seen (%)	Unseen (%)
I3D	Event	Kinetics-400 (R)	55.54	37.52
E^2(GO)-3D	Event	Kinetics-400 (R)	57.87	38.76
TSM	Event	ImageNet	**65.93**	38.23
E^2(GO)-2D	Event	ImageNet	65.40	**40.33**
I3D	Event+RGB	Kinetics-400 (R)	59.12	38.13
E^2(GO)-3D	Event+RGB	Kinetics-400 (R)	61.23	**41.85**
TSM	Event+RGB	ImageNet	71.88	39.92
E^2(GO)-2D	Event+RGB	ImageNet	**72.42**	40.61
I3D	Event+Flow	Kinetics-400 (R)	60.48	44.47
E^2(GO)-3D	Event+Flow	Kinetics-400 (R)	62.66	45.86
TSM	Event+Flow	ImageNet	72.26	46.89
E^2(GO)-2D	Event+Flow	ImageNet	**72.87**	**49.23**
I3D	RGB+Flow	Kinetics-400 (R)	62.07	44.56
TSM	RGB+Flow	ImageNet	**75.08**	**45.66**

Table 3. Accuracy results (%) of the event modality when used in combination to stardard RGB and optical flow. In **bold** the best result for each modality combination.

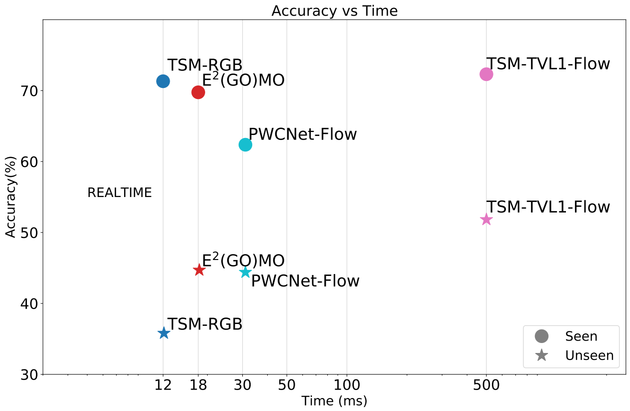

Figure 5. Accuracy *vs* time of RGB modality, E^2(GO)MO, estimated PWCNet optical flow and TV-L1 optical Flow on seen and unseen scenarios for one clip evalutation.

flow. When combined with RGB, it achieves an improvement of up to 7% on seen test sets and 3% on unseen ones. When combing events with optical flow, the best performance is achieved, improving event results by up to 7% on seen domains and 9% on unseen ones. This suggests that, while both event and flow encode motion, flow emphasizes the motion-relevant part, neglecting the scene or object affordances, while the event data maintain useful information about objects' shape (see Figure 3). For this reason, the event modality is potentially more convenient to be combined with optical flow data than with RGB, which, instead, suffer on unseen domains due to its dependency on appearance. It is also worth noting that it outperforms standard RGB+Flow since standard event representations does not emphasize features of appearance as much as RGB does.

E^2(GO)MO. In Table 4 we illustrate the performance of E^2(GO)MO against an RGB-based TSM, which we proved to be the most robust architecture in the previous analysis. To prove our claim that the proposed distillation technique benefits from motion features, we also apply the same mechanism to an RGB-based stream, which we label in Table 4 with the RGB+$\mathcal{L}_{dist}$ entry. Both event and RGB benefit from the flow learning strategy, improving performance on unseen tests (+5.3% and +3% respectively), confirming the importance of motion information in real-world scenarios. However, E^2(GO)MO gains far more from the distillation loss $\mathcal{L}_{dist}$ than RGB, indicating that event data conveys more motion-rich features than RGB streams, thus proving our argument. Finally, we compare these two networks against their multi-modal upper bound performance, obtained exploiting the offline-computed optical flow also in prediction, namely RGB+Flow and E^2(GO)+Flow. Despite both are unable to reach their upper bound, E^2(GO)MO is much closer to E^2(GO)+Flow, and it even exceeds the multi-modal RGB-Flow performance. This result further motivates the use of event data in egocentric vision.

Event *vs*. Optical Flow. We illustrate in Figure 5 the accuracy *vs*. average time per frame trade-off at test time on both seen and unseen data. We report performance with both TV-L1 flow, computed offline [108], and the one ex-

Method	Model	**D1**	**D2**	**D3**	**D1→D2**	**D1→D3**	**D2→D1**	**D2→D3**	**D3→D1**	**D3→D2**	**Seen (%)**	**Unseen (%)**	**Mean (%)**
RGB	TSM	61.61	77.08	75.75	37.39	32.49	34.28	38.99	34.43	38.25	71.48	35.97	53.73
RGB + $\mathcal{L}_{dist}$	TSM	63.36	79.47	77.97	38.61	35.73	39.36	41.09	34.76	49.68	**73.60**	39.87	56.73 ▲+3
RGB + Flow	TSM	66.97	79.69	78.58	43.76	43.76	45.80	47.13	45.44	48.09	75.08	45.66	60.37
Event	TSM	56.86	72.43	68.49	28.73	34.00	37.09	42.30	42.27	45.02	65.93	38.23	52.08
Event	E^2(GO)-2D	56.58	70.03	69.60	34.98	35.16	38.21	47.80	41.71	44.13	65.40	40.33	52.87
Event	E^2(GO)MO-2D	61.38	75.83	75.08	39.77	37.19	44.71	51.03	47.01	53.73	70.76	**45.57**	**58.17 ▲+5.3**
Event + Flow	E^2(GO)-2D	65.11	77.58	75.91	42.12	41.80	48.20	53.50	51.85	57.91	72.87	49.23	61.05

Table 4. Accuracy (%) of E^2(GO)MO w.r.t. the baseline on events (TSM) and E^2(GO)-2D. We compare E^2(GO)MO with the same approach on RGB to validate the choice of combining event and flow. In **bold** the best uni-modal, underlined the best multi-modal.

Stream	**Model**	**Repr. Time (ms)**	**Seen (%)**	**Unseen (%)**
RGB	I3D		**58.49**	35.65
Event	I3D	**6ms**	55.54	37.52
Event	E^2(GO)-3D	**6ms**	57.87	**38.76**
Flow (TV-L1)	I3D	488ms	58.47	43.40
RGB	TSM		**71.48**	35.97
Event	TSM	**6ms**	65.93	38.23
Event	E^2(GO)-2D	**6ms**	65.40	**40.33**
Flow (TV-L1)	TSM	488ms	73.23	53.98

Table 5. Accuracy result of RGB, Event and optical flow (TV-L1), along with their representation time, i.e., time to calculate the Voxel Grid for event, and extraction time for TV-L1 flow.

tracted from PWC-Net [93]. The latter is the most competitive among existing end-to-end CNN models for flow, providing an optimal balance between time and accuracy. For calculation, we use a NVIDIA Titan RTX GPU, and report both input's computation and forward time, ignoring data access time. We also highlight the range under which we can perform real-time action recognition, using the threshold considered in [88] to determine a sufficient frame (sampling) rate for a motion tracking system as a reference point. The plot clearly shows how TV-L1 achieves higher accuracy at the cost of 488 ms of extraction time, making it unsuitable for online scenarios. When the optical flow is estimated online with PWC-Net, performance drops dramatically (by up to 10% on seen tests and 8% on unseen tests). Additionally, PWC-Net necessitates the execution of an additional network, increasing the parameter count (≈ 40M) and requiring an additional fine-tuning stage. In contrast, we do not have to compute flow at test time, thus we can take full advantage of the more precise optical flow when distilling. Despite E^2(GO)MO does not explicitly use flow during inference, it still outperforms PWC-Net on seen tests (by up to 6%) and performs on par with it on unseen ones.

Discussion and Limitations. As it is currently not possible to fully replicate event camera behaviors, event simulation may create undesirable sim-to-real domain shift [79, 91]. Nevertheless, several works showed that simulated events are robust enough to generalize well to real ones [33, 79, 91]. As we introduce event data in egocentric action recognition for the first time, we aim at providing a direct comparison with common benchmarks in the literature [14, 15, 27] and place the event modality in a competitive setting against well-established modalities. These aspects motive us to simulate the event data instead of generating a new first-person dataset from scratch.

Starting from the promising results of our work, we plan to further explore the use of real event streams in this context in order to validate the considerations done so far on a real camera. Moreover, Table 5 shows that, despite its high computational and time cost, TV-L1 optical flow still demonstrates higher performance, especially an extraordinary resiliency to domain changes. We primarily attribute this to the fact that the algorithm for extracting it partially filters out camera motion, resulting in cleaner motion data compared to the unprocessed events. To this purpose, interesting future works could involve the exploitation of motion compensation techniques commonly used with events [90, 90] to remove redundant background noise.

7. Conclusion

In this paper, we presented N-EPIC-Kitchens, the first event-based egocentric action recognition dataset. Exploiting the variety of data modes at our disposal, we carried out an in-depth comparative analysis whose results demonstrate the relevance of motion information in action recognition context. Given these findings, we proposed and evaluated two novel approaches suited for event data (E^2(GO) and E^2(GO)MO) that, by emphasizing motion information, produced competitive results compared to the computational expensive optical flow modality. Through extensive experiments, we bring to light the robustness of event data and their applicability in an online action recognition setting, pushing the community to further explore in this direction.

Acknowledgements. This work was supported by the CINI Consortium through the VIDESEC project. We acknowledge the CINECA award under the ISCRA initiative and IIT HPC infrastructure, for the availability of high performance computing resources and support. A special thanks goes to Antonio Loquercio for his precious advice.

References

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018. 1

[2] Filipp Akopyan, Jun Sawada, Andrew Cassidy, Rodrigo Alvarez-Icaza, John Arthur, Paul Merolla, Nabil Imam, Yutaka Nakamura, Pallab Datta, Gi-Joon Nam, et al. Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip. *IEEE transactions on computer-aided design of integrated circuits and systems*, 34(10):1537–1557, 2015. 2

[3] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7243–7252, 2017. 2, 3

[4] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based spatio-temporal feature learning for neuromorphic vision sensing. *IEEE Transactions on Image Processing*, 29:9084–9098, 2020. 2

[5] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. Asynchronous convolutional networks for object detection in neuromorphic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 4

[6] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *European Conference on Computer Vision*, pages 136–152. Springer, 2020. 4

[7] Marco Cannici, Chiara Plizzari, Mirco Planamente, Marco Ciccone, Andrea Bottino, Barbara Caputo, and Matteo Matteucci. N-rod: A neuromorphic dataset for synthetic-to-real domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1342–1347, 2021. 2

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 4, 6

[9] Alejandro Cartas, Jordi Luque, Petia Radeva, Carlos Segura, and Mariella Dimiccoli. Seeing and hearing egocentric actions: How much can we learn? In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[10] Junming Chen, Jingjing Meng, Xinchao Wang, and Junsong Yuan. Dynamic graph cnn for event-camera based gesture recognition. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2020. 2

[11] Gregory Kevin Cohen. *Event-Based Feature Detection, Recognition and Classification*. Theses, Université Pierre et Marie Curie - Paris VI ; University of Western Sydney, Sept. 2016. 4

[12] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019. 2, 3

[13] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021. 6

[14] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1, 2, 3, 4, 8

[15] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2021. 1, 3, 4, 8

[16] Dima Damen, Adriano Fragomeni, Jonathan Munro, Toby Perrett, Daniel Whettam, and Michael Wray. Epic-kitchens-100- 2021 challenges report. https://epic-kitchens.github.io/Reports/EPIC-KITCHENS-Challenges-2021-Report.pdf, 2021. 2, 3

[17] Dima Damen, Evangelos Kazakos, Will Price, Jian Ma, and Hazel Doughty. Epic-kitchens-55 - 2020 challenges report. https://epic-kitchens.github.io/Reports/EPIC-KITCHENS-Challenges-2020-Report.pdf, 2020. 2, 3

[18] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio W Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, volume 2, page 3, 2014. 1

[19] Dima Damen, Will Price, Evangelos Kazakos, Antonino Furnari, and Giovanni Maria Farinella. Epic-kitchens - 2019 challenges report. https://epic-kitchens.github.io/Reports/EPIC-Kitchens-Challenges-2019-Report.pdf, 2019. 2, 3

[20] Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010. 1

[21] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *Ieee Micro*, 38(1):82–99, 2018. 2

[22] Pierre de Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*, 2020. 3

[23] Ana Garcia Del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47(1):65–76, 2016. 1

[24] Tobi Delbruck. Neuromorophic vision sensing and processing. In *2016 46Th european solid-state device research conference (ESSDERC)*, pages 7–14. IEEE, 2016. 2

[25] 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. 4

[26] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[27] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pages 314–327. Springer, 2012. 3, 8

[28] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020. 2

[29] Steve B Furber, David R Lester, Luis A Plana, Jim D Garside, Eustace Painkras, Steve Temple, and Andrew D Brown. Overview of the spinnaker system architecture. *IEEE Transactions on Computers*, 62(12):2454–2467, 2012. 2

[30] Antonino Furnari and Giovanni Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 3

[31] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *arXiv preprint arXiv:1904.08405*, 2019. 2

[32] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 3

[33] 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. 4, 8

[34] 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. 4

[35] Daniel Gehrig, Michelle Ruegg, Mathias Gehrig, Javier Hidalgo-Carrio, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, apr 2021. 2

[36] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 3

[37] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *International Conference on 3D Vision (3DV)*, 2021. 2

[38] 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. 6

[39] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. *arXiv preprint arXiv:2106.02036*, 2021. 1

[40] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. *arXiv preprint arXiv:1711.01467*, 2017. 2

[41] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*, 2021. 1, 3

[42] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 6

[43] 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. 2, 6

[44] Javier Hidalgo-Carrio, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. In *2020 International Conference on 3D Vision (3DV)*. IEEE, nov 2020. 2

[45] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[46] Yuhuang Hu, Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE, 2020. 2

[47] 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. 3

[48] Simone Undri Innocenti, Federico Becattini, Federico Pernici, and Alberto Del Bimbo. Temporal binary representation for event-based action recognition. *arXiv*, 2020. 2, 3, 4

[49] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal co-

variate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 2, 6

[50] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018. 3

[51] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. *arXiv preprint arXiv:2111.01024*, 2021. 3

[52] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 1, 3, 5

[53] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13618–13627, 2021. 1, 3

[54] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2146–2156, 2021. 3, 6

[55] 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 transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2016. 4

[56] Myunggi Lee, Seungeui Lee, Sungjoon Son, Gyutae Park, and Nojun Kwak. Motion feature network: Fixed motion filter for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 387–403, 2018. 3

[57] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012. 1

[58] Yong Jae Lee and Kristen Grauman. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision*, 114(1):38–55, 2015. 1

[59] Yin Li, Miao Liu, and Jame Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[60] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015. 3

[61] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 166:41–50, 2018. 2

[62] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 2, 4, 5, 6

[63] Yihan Lin, Wei Ding, Shaohua Qiang, Lei Deng, and Guoqi Li. Es-imagenet: A million event-stream classification dataset for spiking neural networks. *arXiv preprint arXiv:2110.12211*, 2021. 3

[64] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *European Conference on Computer Vision*, pages 704–721. Springer, 2020. 1

[65] Yao Lu and Walterio W Mayol-Cuevas. Understanding egocentric hand-object interactions from hand pose estimation. *arXiv preprint arXiv:2109.14657*, 2021. 1

[66] Iulia-Alexandra Lungu, Federico Corradi, and Tobi Delbrück. Live demonstration: Convolutional neural network driven by dynamic vision sensor playing roshambo. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–1. IEEE, 2017. 3

[67] Iulia Alexandra Lungu, Shih-Chii Liu, and Tobi Delbruck. Incremental learning of hand symbols using event-based cameras. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(4):690–696, 2019. 2

[68] Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997. 2, 4

[69] Nico Messikommer, Daniel Gehrig, Antonio Loquercio, and Davide Scaramuzza. Event-based asynchronous sparse convolutional networks. In *European Conference on Computer Vision*, pages 415–431. Springer, 2020. 2

[70] Shu Miao, Guang Chen, Xiangyu Ning, Yang Zi, Kejia Ren, Zhenshan Bing, and Alois Knoll. Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. *Frontiers in neurorobotics*, 13:38, 2019. 2, 3

[71] Kyle Min and Jason J Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1069–1078, 2021. 3

[72] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 122–132, 2020. 1, 2, 3, 6

[73] Chethan M Parameshwara, Nitin J Sanket, Chahat Deep Singh, Cornelia Fermüller, and Yiannis Aloimonos. 0-mms: Zero-shot multi-motion segmentation with a monocular event camera. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9594–9600. IEEE, 2021. 2

[74] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An

imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[75] Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *arXiv preprint arXiv:2009.13436*, 2020. 3

[76] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021. 6

[77] Mirco Planamente, Andrea Bottino, and Barbara Caputo. Self-supervised joint encoding of motion and appearance for first person action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8751–8758. IEEE, 2021. 3

[78] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Domain generalization through audio-visual relative norm alignment in first person action recognition. *arXiv preprint arXiv:2110.10101*, 2021. 1, 3

[79] 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 preprint arXiv:2103.12768*, 2021. 2, 4, 8

[80] Will Price and Dima Damen. An evaluation of action recognition models on epic-kitchens. *arXiv preprint arXiv:1908.00867*, 2019. 4

[81] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999. 6

[82] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 2

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

[84] Ivan Rodin, Antonino Furnari, Dimitrios Mavroeidis, and Giovanni Maria Farinella. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 211:103252, 2021. 3

[85] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. *arXiv preprint arXiv:2110.15128*, 2021. 1, 3

[86] Yusuke Sekikawa, Kosuke Hara, and Hideo Saito. Eventnet: Asynchronous recursive event processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3887–3896, 2019. 2

[87] 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. 4

[88] Min-Ho Song and Rolf Inge Godøy. How fast is your body motion? determining a sufficient frame rate for an optical motion tracking system using passive markers. *PloS one*, 11(3):e0150993, 2016. 8

[89] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9787–9795, 2021. 1, 3

[90] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7244–7253, 2019. 8

[91] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 534–549. Springer, 2020. 4, 8

[92] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9954–9963, 2019. 2, 3

[93] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 8

[94] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4597–4605, 2015. 2

[95] Daksh Thapar, Aditya Nigam, and Chetan Arora. Anonymizing egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2320–2329, 2021. 1

[96] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[97] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019. 2

[98] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[99] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide

Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16155–16164, 2021. 2

[100] Ajay Vasudevan, Pablo Negri, Bernabe Linares-Barranco, and Teresa Serrano-Gotarredona. Introduction and analysis of an event-based sign language dataset. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 675–682. IEEE, 2020. 3

[101] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4006–4015, 2019. 3

[102] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 2, 6

[103] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 2, 4, 6

[104] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8168–8177, 2021. 3

[105] Yangming Wen, Krishna Kumar Singh, Markham Anderson, Wei-Pang Jan, and Yong Jae Lee. Seeing the unseen: Predicting the first-person camera wearer's location and pose in third-person scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3446–3455, 2021. 1

[106] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 2

[107] Jiancheng Yang, Qiang Zhang, Bingbing Ni, Linguo Li, Jinxian Liu, Mengdie Zhou, and Qi Tian. Modeling point clouds with self-attention and gumbel subset sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3323–3332, 2019. 2

[108] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007. 2, 7

[109] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021. 6

[110] Jiaojiao Zhao and Cees GM Snoek. Dance with flow: Two-in-one stream action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9935–9944, 2019. 3

[111] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 2

[112] Yi Zhou, Guillermo Gallego, Xiuyuan Lu, Siqi Liu, and Shaojie Shen. Event-based motion segmentation with spatio-temporal graph cuts. *IEEE Transactions on Neural Network and Learning Systems*, 2021. 2

[113] 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, 4