2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Event-based Few-shot Fine-grained Human Action Recognition

Zonglin Yang[1], Yan Yang[2], Yuheng Shi[1],
Hao Yang[1], Ruikun Zhang[1], Liu Liu[3], Xinxiao Wu[1] and Liyuan Pan[1]

***Abstract*— Few-shot fine-grained human (FGH) action recognition is crucial in the context of human-robot interaction within open-set real-world environments. Existing works mainly focus on features extracted from RGB frames. However, their performances are drastically impacted in challenging scenarios, such as high-dynamic or low lighting conditions. Event cameras can independently and sparsely capture brightness changes in a scene at microsecond resolution and high dynamic range, which offer a promising solution. However, the modality differences between events and RGB frames, and the lack of paired fine-grained data hinder the development of event-based FGH action recognition. Therefore, in this paper, we introduce the first Event Camera Fine-grained Human Action (E-FAction) dataset. This dataset comprises 3304 paired 'event stream and RGB sequence', covering 15 coarse action classes and 128 fine-grained actions. Then, we develop a versatile event feature extractor. Considering the spatial sparsity of event stream, we design two modules to mine the temporal motion and semantic features under the guidance of paired RGB frames, facilitating robust weight initialization for the feature extractor in few-shot FGH action recognition. We conduct extensive experiments on both published and our built synthetic and real datasets, and consistently achieve state-of-the-art performance compared to existing baselines. Code and dataset will be available at *link*.**

I. INTRODUCTION

For the human-robot interaction field, robots should accurately recognize and adapt to a wide range of fine-grained human (FGH) actions in real-world settings [1]. Few-shot classification has been adopted in this field to classify unseen fine-grained human actions from a small set of examples [2], [3]. Compared to traditional few-shot FGH action recognition approaches using RGB frames, employing stream of events (*e.g.*, event stream [4]) is more appealing due to its robustness in low-light conditions or motion blur, and its potential to alleviate privacy concerns to some extent [5], [6], [7].

Event cameras are bio-inspired sensors [8], [4] that generate event stream by asynchronously and sparsely recording the scene brightness changes with microsecond temporal resolution and high dynamic range (120dB vs 60dB of RGB cameras) [9]. It benefits the FGH action recognition task in two aspects: i) the event stream robustly captures subtle human movements at high temporal resolution [1]. This enables models to more effectively interpret the temporal relationships that are essential for recognizing actions; ii) the event stream is robust to motion blur and low lighting conditions compared to RGB frames [4]. This enables the model to handle challenging scenarios with fast motion and high-contrast scenes. Therefore, in this paper, we aim to design an event-based few-shot FGH action recognition approach which is accurate in challenging scenarios. To the authors' best knowledge, this paper is the first work that explores the potential of event stream on challenging scenarios for the FGH action recognition task.

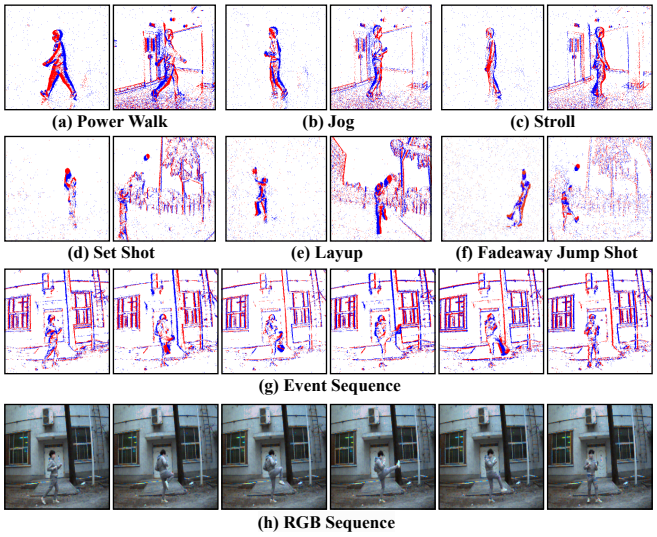


Fig. 1. Examples of our proposed E-FAction dataset. (a)-(c) and (d)-(f) show the event stream for example actions from coarse class 'Gait' and 'Basketball' separately, with fine-grained human action annotations. Red and blue pixels depict positive and negative events, respectively. (g) and (h) show an example of a paired event sequence and RGB sequence of 'Boxing - Front Kick'. Best viewed in colour on the screen.

However, the absence of datasets hinders the development of event-based few-shot FGH action recognition algorithms. Hence, in this paper, we introduce an Event Camera Fine-grained Human Action (E-FAction) dataset, containing 3304 event streams and their paired RGB sequences collected under a real-world setting, categorized into 15 coarse and 128 fine-grained human actions. Refer to Fig. 1 for examples of our dataset. Our E-FAction dataset stands as the first event-based benchmark dataset, which provides tri-modality data, *i.e.*, event stream, RGB frames, and CoCo-style captions, on various challenging real-world scenarios.

With our collected dataset, one may directly use existing RGB-based algorithms by generally fine-tuning their feature extractors [2], [10]. However, it can lead to suboptimal performance due to the domain divergence between RGB and event domain (refer to experiment results in Sec. V-C)[11]. For example, RGB frames record scene brightness in a dense, uniform manner, whereas event stream records changes in brightness and is spatially sparse. To this end, we propose a

[1] School of CSAT, Beijing Institute of Technology.
[2] BDSI, College of Science, The Australian National University.
[3] KooMap Dept., Huawei.

new feature extractor for event stream to seamlessly leverage the existing RGB-based algorithms for event-based few-shot FGH action recognition.

To efficiently learn the event feature extractor, we guide the feature extraction process using its paired RGB frames and pre-trained RGB feature extractors. Considering that our few-shot action recognition task requires reasoning about temporal motions and semantics [12], we propose two modules for enforcing feature consistency between event stream and RGB frames, using the pre-trained CLIP image encoder [13] and VideoFlow motion encoder [14], respectively. Two optimization objectives are used for providing a robust weight initialization for the event feature extractor in the following few-shot action recognition, improving the recognition performance. For example, combining our event feature extractor with the state-of-the-art few-shot action recognition algorithm from [15] finds 84.8% accuracy on $\mathrm{THU}^{\text{E-ACT}}_{-50}$ dataset [16], which is 5.6% higher than using their default encoder.

Our contributions are summarized into three folds:

- We present the first fine-grained human action recognition dataset for the event camera, E-FAction, which contains 128 fine-grained human actions on various challenging scenarios with paired event stream, RGB sequences, and captions.
- We introduce a versatile event feature extractor by enforcing consistency of temporal motions and semantic features between event stream and RGB frames. The learned extractor seamlessly incorporates existing algorithms for few-shot FGH action recognition.
- We design the first settings for few-shot FGH action recognition, benchmark extensive baseline methods, and consistently achieve state-of-the-art performance using our event feature extractor on synthetic and real event datasets.

II. RELATED WORK

We introduce recent studies on event-based human action recognition, and then review event-based human action datasets.

A. Fine-grained Event-based Human Action Recognition

Recent progress in event-based fine-grained human action recognition has been achieved from two main perspectives: event representation and network architecture. i) Event representation involves converting event stream into 2D grids, known as event frames, using either fixed [17], [18], [19] or learnable methods [1], [20] to adapt them for use with networks designed for RGB frames. ii) Networks [21], [22], [23] are designed to address challenges of spatial sparsity and temporal dynamics in event frames for action recognition. Other approaches study graph neural networks [21] and spiking neural networks [24], [25] to work on the event stream without the event frame conversion. However, for network performance, they use extensive labeled event stream in training, and are hard to be adapted in open-set real-world environments.

For real-world applications, few-shot fine-grained action recognition has been explored that learns a hypersurface effectively for recognizing human actions from a small set of labeled examples, known as a support set. Recent strategies leverage meta-learning and vision-language model fine-tuning. i) Meta-learning trains model to quickly adapt to new tasks with a small support set by employing a 'learning to learn' strategy [26]. It can use memory structures to cache label-related features for augmenting input features [27], dynamic temporal alignment algorithms [28] to align temporal information of an input with those in the support set, and attention layers to fuse spatio-temporal features from the support set for an input [10]. ii) Vision-language fine-tuning uses large-scale pre-trained vision-language models. It adds new layers to the vision-language models and fine-tunes them for few-shot action recognition [15], [29]. Compared to the meta-learning method, the rich vision and language knowledge embedded within these models is more helpful in understanding and classifying fine-grained actions with minimal support sets. Despite extensive research in few-shot action recognition using RGB cameras, the application of event cameras in this field has not been explored. We propose a versatile event feature extractor to use the vision-language-based methods for few-shot fine-grained action recognition.

B. Event-based Human Action Datasets

The landscape of datasets in event-based action recognition is continuously evolving. Earlier contributions such as DHP19 [33] and DailyAction [34] datasets lay the groundwork, albeit with limited action diversity and dataset size. To avoid the limitations, HMDB-DVS [31], [35] and UCF-DVS [31], [36] datasets are converted from RGB sequences, borrowing extensive repository of human action datasets in the RGB domain. With growing interest in human action recognition for event stream, recent works collect large-scale and diverse data from real-world environments, such as $\mathrm{THU}^{\mathrm{E-ACT}}_{-50}$ [22] dataset, which is a significant step forward in terms of scale and diversity for real event datasets. Nonetheless, none of the datasets is designed for fine-grained human action recognition that studies subtly different human actions. This paper introduces E-FAction, the first dataset for fine-grained action recognition using event stream. E-FAction aims to set a new benchmark in the domain.

III. THE E-FACTION DATASET: A TRI-MODAL BENCHMARK

A. Method

We craft the E-FAction (Event Camera Fine-grained Human Action) dataset for benchmarking event-based FGH action recognition by following principles.

Tri-modality. Our dataset comprises event stream, RGB frames with precise timestamps, and captions. The captions are crafted for event stream within each interval of adjacent RGB frames, and are in a similar style to the CoCo dataset [37] by two human experts with assistance from LLaVA [38]. The examples of these three modalities are displayed in Fig. 3.

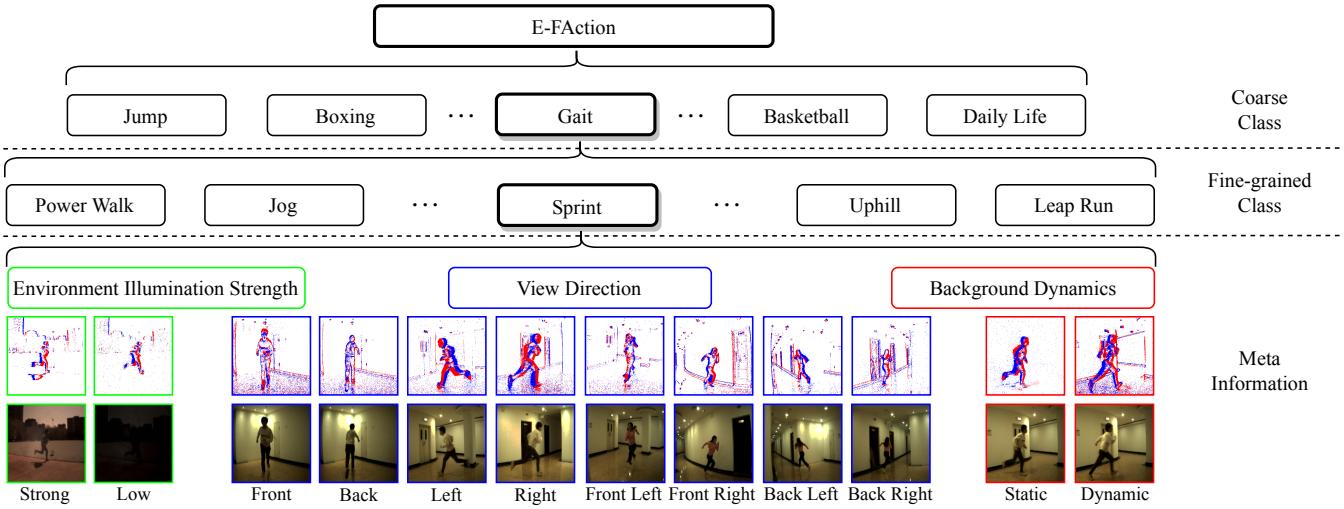


Fig. 2. The human action hierarchy of our E-FAction (Event Camera Fine-grained Human Action) dataset. It has 15 coarse and 128 fine-grained action classes, with meta information of View Direction, Environment Illumination Strength, and Background Dynamics. We take coarse class 'Gait' as an example to illustrate the fine-grained classes that belong to it. We further display the three types of meta information for 'Sprint' using distinct colors, each comprising both RGB frames and event stream. Red and blue pixels depict positive and negative events, respectively. The Environment Illumination Strength includes both strong and low lighting conditions. In terms of the View Direction, it encompasses front, back, left, right, and diagonal directions. The Background Dynamics is categorized into pure human action (static) and scenes that include the complete background (dynamic). Best viewed in colour on the screen.

TABLE I

COMPARISON OF EVENT DATASETS FOR HUMAN ACTION RECOGNITION. THE RGB, CAPTION, MI, MV, DB, AND I/O DENOTES WHETHER THE RGB FRAMES ARE PUBLICLY AVAILABLE, FRAME-WISE PRECISE CAPTION, MULTI-ILLUMINATION, MULTI-VIEW, DYNAMIC BACKGROUND, AND WHETHER BOTH INDOOR AND OUTDOOR SCENES ARE AVAILABLE, RESPECTIVELY.

Dataset	Year	Sensor	Fine-grained	RGB	Caption	MI	MV	DB	I/O	Real	Scale	Class	Resolution
DvsGesture [30]	2017	DAVIS128	✗	✗	✗	✓	✗	✗	✗	✓	1,342	11	128×128
HMDB-DVS [31]	2019	DAVIS240c	✗	✗	✗	✗	✗	✗	✗	✗	6,766	51	240×180
PAF [32]	2019	DAVIS346	✗	✗	✗	✗	✗	✗	✗	✓	450	10	346×260
DHP-19 [33]	2019	DAVIS346	✗	✗	✗	✗	✓	✗	✗	✓	561	33	346×260
DailyAction [34]	2021	DAVIS128	✗	✗	✗	✗	✗	✗	✗	✓	1,440	12	128×128
$\text{THU}^{\text{E-ACT}}_{-50}$ [16]	2023	CeleX-V	✗	✗	✗	✗	✓	✗	✓	✓	10,500	50	1280×800
E-FAction	2024	DAVIS346	✓	✓	✓	✓	✓	✓	✓	✓	3,304	128	346×260

Various Challenges. Our dataset is collected from real-world scenarios, and incorporates diverse challenging factors: i) Multi-illumination. To explore the potential of high dynamic range in an event camera, we record human actions under low and strong light environments; ii) Multi-view. We collect data from eight distinct views to the human actor, front, back, left, right, and four diagonal directions; iii) Diverse scenes. The dataset mimics the variability of real-world environments by encompassing indoor and outdoor scenes. For example, corridors, offices, playgrounds, and squares; iv) Dynamic background. To reflect extensive application scenarios of FGH action recognition, our dataset involves static and dynamic backgrounds for human actions; v) Action duration. In the real world, the speed of the same human action may be different. For our dataset, we set the duration of FGH actions varying from 2 to 8 seconds; vi) Hierarchical categorization. The human action classes of the E-FAction dataset are organized into a hierarchical structure, starting with 15 coarse classes to 128 fine-grained classes. The hierarchy of the classes of human actions is illustrated in Fig. 2, with samples of the dataset presented. We will release our dataset publicly with the meta information above.

B. Collection Methodology and Dataset Division

The dataset uses the DAVIS346 event camera [8] to capture 3304 action sequences in a resolution of 346×260 from real-world scenarios. The E-FAction dataset provides a challenging benchmark for recognizing fine-grained human actions, such as differentiating between 'Power Walk', 'Jog', and 'Stroll'. All collection processes adhere to the aforementioned protocols, carried out by 10 action performers. The dataset is divided into training (60%), validation (15%), and testing (25%) sets, facilitating the development and evaluation of future event-based human action recognition models. We compare the E-FAction dataset with existing action classification datasets in Tab. I.

Recognizing the sensitive nature of our dataset, all data involving human participants were collected following informed consent procedures. Participants were fully informed about the study's aims, the nature of their involvement,

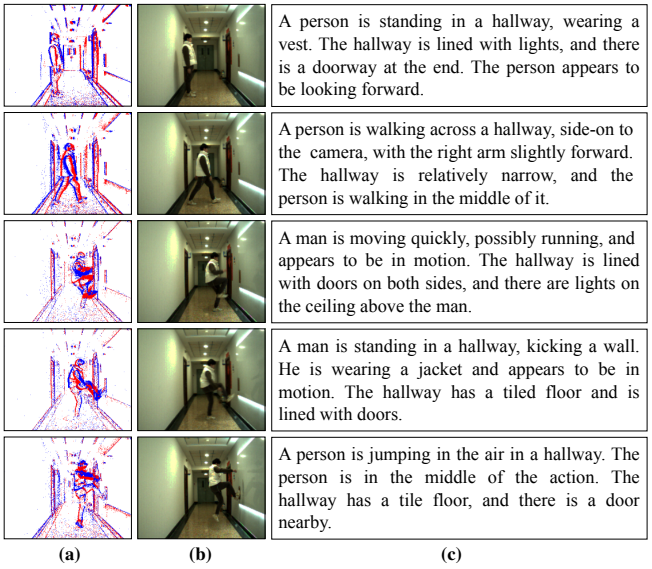


Fig. 3. Examples of the tri-modality data of action 'Jump - Wall Kick' from the E-FAction dataset. (a)-(c) show the event stream, RGB frames, and captions from left to right. Best viewed in colour on the screen.

potential risks, benefits, and their right to withdraw from the study at any point without any consequences.

IV. FEW-SHOT FINE-GRAINED HUMAN ACTION RECOGNITION

Given event stream $\mathcal{E} = \{x_i, y_i, s_i, p_i\}_{i=1}^{N}$ of size N and T paired RGB frames $\{\mathcal{X}_i^{\text{rgb}}\}_{i=1}^{T}$, our goal is to learn a versatile event feature extractor $\mathbf{E}^{\text{evt}}(\cdot)$ for few-shot FGH action recognition, where (x_i, y_i), s_i, and p_i are position, timestamp, and polarity of the i-th event.

Our method has three key steps: i) RGB and event alignment, converting event stream to T event frames $\{\mathcal{X}_i^{\text{evt}}\}_{i=1}^{T}$ by aligning event stream to corresponding time slots of the RGB frames; ii) motion and semantic feature learning, using a pre-trained CLIP [13] image encoder $\mathbf{E}^{\text{clip}}(\cdot)$ and a pre-trained VideoFlow [14] encoder $\mathbf{E}^{\text{vif}}(\cdot)$ to extract RGB frame features for supervising our event feature extractor $\mathbf{E}^{\text{evt}}(\cdot)$, with our semantic and motion feature learning losses, $\mathcal{L}_{\text{m}}$ and $\mathcal{L}_{\text{s}}$; iii) few-shot fine-grained action recognition, using our learned event feature extractor with existing few-shot algorithms for FGH action recognition. The overview of our framework is shown in Fig. 4.

A. RGB and Event Alignment

We convert the asynchronous event stream into synchronous event frames based on the exposure time Δs of the RGB frames $\{\mathcal{X}_i^{\text{rgb}}\}_{i=1}^{T}$ for better alignment. Note that the time span of the event stream $\mathcal{E}$ is ΔsT. We first use the event representation [39] for converting $\mathcal{E}$ to B-channel event volumes $\mathcal{V}$ by

$$\mathbf{R}(s) = (B-1)\frac{s - s_1}{\Delta sT}\ , \quad \mathbf{K}_b(a) = \max(0, 1 - |a|)\ ,$$

$$\mathcal{V}(x, y, b) = \sum_{i=1}^{N} |p_i| \mathbf{K}_b(1 - x_i)\mathbf{K}_b(1 - y_i)\mathbf{K}_b(1 - \mathbf{R}(s_i))\ ,$$

where $\mathbf{R}(s)$ normalizes the input timestamp s, $\mathbf{K}_b(a)$ is a bilinear interpolation function with variable a, and $\mathcal{V}(x, y, b)$ is the value at the x-th row, y-th column, and b-th channel. Then, we uniformly split the event volume $\mathcal{V}$ into T continuous chunks, finding our event frames $\{\mathcal{X}_i^{\text{evt}}\}_{i=1}^{T}$.

B. Motion and Semantic Feature Learning

Motion Learning Module. For FGH action recognition, the temporal motions should be extracted for accurately predicting the action class [12]. We enforce our event feature extractor $\mathbf{E}^{\text{evt}}(\cdot)$ to be consistent with motion features extracted by the pre-trained VideoFlow encoder $\mathbf{E}^{\text{vif}}(\cdot)$.

We forward the RGB frames $\{\mathcal{X}_i^{\text{rgb}}\}_{i=1}^{T}$ to the VideoFlow encoder $\mathbf{E}^{\text{vif}}(\cdot)$ by $\mathbf{E}^{\text{vif}}(\{\mathcal{X}_i^{\text{rgb}}\}_{i=1}^{T})$, and apply an average pooling function to collapse the spatial dimension of the VideoFlow encoder output. The averaged features are our RGB frame motion features $\{\mathcal{M}_i^{\text{rgb}}\}_{i=1}^{T}$, and $\mathcal{M}_i^{\text{rgb}} \in \mathbb{R}^{D^{\text{vif}}}$ is a D^{vif}-dimensional vector.

Meanwhile, we embed the event frames $\{\mathcal{X}_i^{\text{evt}}\}_{i=1}^{T}$ by our event feature extractor $\mathbf{E}^{\text{evt}}(\cdot)$, and predict the motion features from the extracted embeddings with a task head $\mathbf{H}^{\text{vif}}(\cdot)$, while collapsing the spatial dimension and unifying the feature dimension to D^{vif}. Similarly, we denote the predicted motion features as $\{\mathcal{M}_i^{\text{evt}}\}_{i=1}^{T}$, with $\mathcal{M}_i^{\text{evt}} \in \mathbb{R}^{D^{\text{vif}}}$.

With the motion features $\{\mathcal{M}_i^{\text{rgb}}\}_{i=1}^{T}$ predicted from RGB frames and $\{\mathcal{M}_i^{\text{evt}}\}_{i=1}^{T}$ that from event frames, we are ready to supervise our event feature extractor for learning motion features. We normalize the motion features $\{\mathcal{M}_i^{\text{rgb}}\}_{i=1}^{T}$ and $\{\mathcal{M}_i^{\text{evt}}\}_{i=1}^{T}$ with a Softmax function into $\{\hat{\mathcal{M}}_i^{\text{rgb}}\}_{i=1}^{T}$ and $\{\hat{\mathcal{M}}_i^{\text{evt}}\}_{i=1}^{T}$, with

$$\hat{\mathcal{M}}_i(j) = \frac{\exp\left(\mathcal{M}_i(j)\right)}{\sum_{j'=1}^{D^{\text{vif}}} \exp\left(\mathcal{M}_i(j')\right)}\ , \tag{1}$$

where $\hat{\mathcal{M}}_i(j)$ can be either $\mathcal{M}_i^{\text{rgb}}$ or $\mathcal{M}_i^{\text{evt}}$, and $\hat{\mathcal{M}}_i(j)$ is the j-th feature value. We then use a Kullback-Leibler (KL) divergence to penalize the deviations of $\{\hat{\mathcal{M}}_i^{\text{rgb}}\}_{i=1}^{T}$ from $\{\hat{\mathcal{M}}_i^{\text{evt}}\}_{i=1}^{T}$, which is our temporal motion loss $\mathcal{L}_{\text{m}}$,

$$\mathcal{L}_{\text{m}} = \frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{D^{\text{vif}}} -\mathcal{M}_i^{\text{rgb}}(j) \log \frac{\mathcal{M}_i^{\text{evt}}(j)}{\mathcal{M}_i^{\text{rgb}}(j)}\ . \tag{2}$$

Semantic Learning Module. We learn the semantics of event frames $\{\mathcal{X}_i^{\text{evt}}\}_{i=1}^{T}$ to help the FGH action recognition. The semantic features of RGB and event frames, $\{\mathcal{S}_i^{\text{rgb}}\}_{i=1}^{T}$ and $\{\mathcal{S}_i^{\text{evt}}\}_{i=1}^{T}$ are extracted by

$$\{\mathcal{S}_i^{\text{rgb}}\}_{i=1}^{T} = \{\mathbf{E}^{\text{clip}}(\mathcal{X}_i^{\text{rgb}})\}_{i=1}^{T}\ , \tag{3}$$

$$\{\mathcal{S}_i^{\text{evt}}\}_{i=1}^{T} = \mathbf{H}^{\text{clip}}\big(\mathbf{E}^{\text{evt}}(\{\mathcal{X}_i^{\text{evt}}\}_{i=1}^{T})\big)\ , \tag{4}$$

where $\mathcal{S}_i^{\text{rgb}} \in \mathbb{R}^{D^{\text{clip}}}$, $\mathcal{S}_i^{\text{evt}} \in \mathbb{R}^{D^{\text{clip}}}$, D^{clip} is the semantic feature dimension, and $\mathbf{H}^{\text{clip}}(\cdot)$ is a task head.

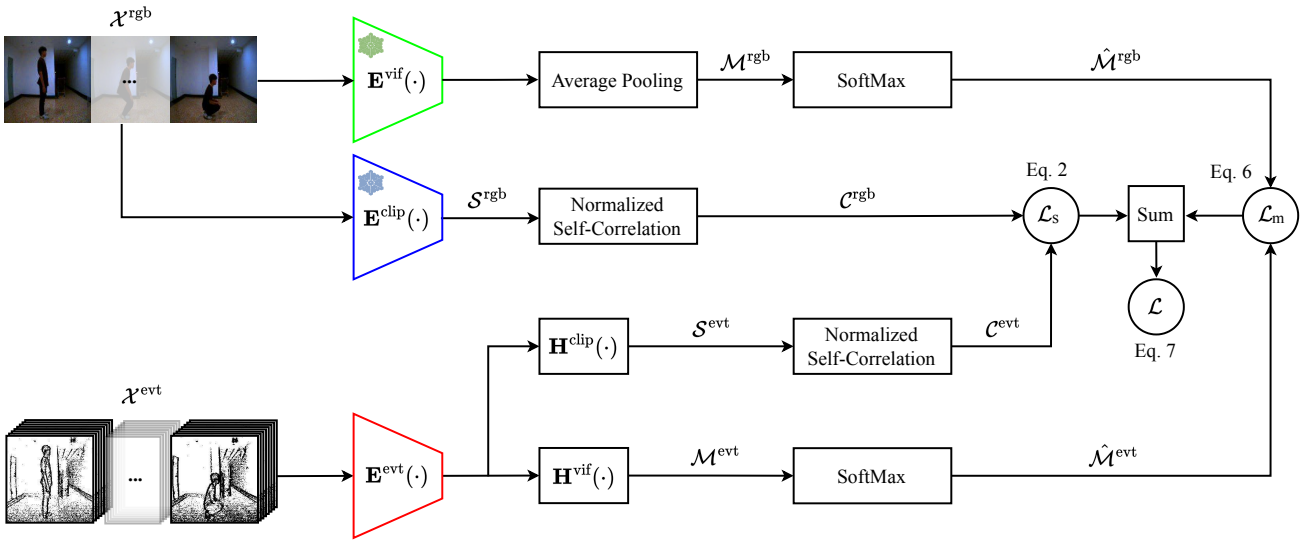


Fig. 4. An overview of our framework for learning an event feature extractor. Our event feature extractor $\mathbf{E}^{\text{evt}}(\cdot)$ is trained to learn motion and semantic features of event frames $\mathcal{X}^{\text{evt}}$. i) We use a pre-trained VideoFlow encoder $\mathbf{E}^{\text{vif}}(\cdot)$ to extract normalized motion features $\hat{\mathcal{M}}^{\text{rgb}}$ of paired RGB frames $\mathcal{X}^{\text{rgb}}$ with using average pooling over spatial dimension and SoftMax over the feature dimension. We predict the motion features with $\mathbf{E}^{\text{evt}}(\cdot)$ and a task head $\mathbf{H}^{\text{clip}}(\cdot)$ of $\mathcal{X}^{\text{evt}}$, and normalize the output $\mathcal{M}^{\text{evt}}$ into $\hat{\mathcal{M}}^{\text{evt}}$ by SoftMax. ii) We extract the semantic features $\mathcal{S}^{\text{rgb}}$ of $\mathcal{X}^{\text{rgb}}$ with a pre-trained CLIP image encoder, and predict semantic features $\mathcal{S}^{\text{evt}}$ of $\mathcal{X}^{\text{evt}}$ with $\mathbf{E}^{\text{evt}}(\cdot)$ and a task head $\mathbf{H}^{\text{clip}}(\cdot)$. The normalized temporal semantic correlation in $\mathcal{S}^{\text{rgb}}$ and $\mathcal{S}^{\text{evt}}$ are calculated, which are $\mathbf{C}^{\text{rgb}}$ and $\mathbf{C}^{\text{evt}}$. iii) We train $\mathbf{E}^{\text{evt}}(\cdot)$ by using a temporal motion loss $\mathcal{L}_{\text{m}}$ and a temporal semantic loss $\mathcal{L}_{\text{s}}$ to supervise $\hat{\mathcal{M}}^{\text{evt}}$ and $\mathbf{C}^{\text{evt}}$ with $\hat{\mathcal{M}}^{\text{rgb}}$ and $\mathbf{C}^{\text{rgb}}$. Best viewed in colour on the screen.

Directly enforcing the feature consistency between $\{\mathcal{S}_i^{\text{rgb}}\}_{i=1}^T$ and $\{\mathcal{S}_i^{\text{evt}}\}_{i=1}^T$ is suboptimal. Considering that FGH action recognition uses semantic variations along the temporal dimension to infer a fine-grained class [40], we calculate the normalized temporal semantic feature correlations $\mathcal{C}^{\text{rgb}}$ and $\mathcal{C}^{\text{evt}}$ for $\{\mathcal{S}_i^{\text{rgb}}\}_{i=1}^T$ and $\{\mathcal{S}_i^{\text{evt}}\}_{i=1}^T$ separately,

$$\mathcal{C}(i,j) = \frac{\exp\left(\mathcal{S}_i \odot \mathcal{S}_j\right)}{\sum_{j'=1}^{T} \exp\left(\mathcal{S}_i \odot \mathcal{S}_{j'}\right)} , \tag{5}$$

where $\mathcal{C}$ is either $\mathcal{C}^{\text{rgb}}$ or $\mathcal{C}^{\text{evt}}$, $(i,j) \in \{1,\cdots,T\} \times \{1,\cdots,T\}$, the value at i-th row and j-th column of $\mathcal{C}$ is $\mathcal{C}(i,j)$, and $\odot$ is a scalar product operation. We then enforce the semantic consistency with our temporal semantic loss $\mathcal{L}_{\text{s}}$,

$$\mathcal{L}_{\text{s}} = \sum_{i=1}^{T}\sum_{j=1}^{T} -\mathcal{C}^{\text{rgb}}(i,j) \log \frac{\mathcal{C}^{\text{evt}}(i,j)}{\mathcal{C}^{\text{rgb}}(i,j)} . \tag{6}$$

Overall Loss. Our network is optimized by $\mathcal{L}$ which combines the temporal motion loss $\mathcal{L}_{\text{m}}$ and the temporal semantic loss $\mathcal{L}_{\text{s}}$. Balanced by a hyperparameter λ, the loss $\mathcal{L}$ is

$$\mathcal{L} = \lambda\mathcal{L}_{\text{m}} + \mathcal{L}_{\text{s}} . \tag{7}$$

C. Few-shot Fine-grained Human Action Recognition

We follow the state-of-the-art few-shot FGH action recognition method [15], substituting the default image encoder with our event feature extractor. Specifically, we first use a CLIP text encoder [13] to extract action semantics from the text of actions, then refine extracted event features with a feedforward network, and calculate semantic similarity between the extracted text features and refined event features to classify a FGH action. We also combine our event feature extractor with other few-shot FGH action recognition methods in the experiment sections to demonstrate the effectiveness of our extractor.

V. EXPERIMENTS

A. Datasets

We evaluate our proposed framework on four datasets: N-FineGym [12], E-FAction, DailyAction [34], and $\text{THU}^{\text{E-ACT}}_{-50}$ [16] for event-based few-shot action recognition. Considering the lack of existing event-based FGH action recognition datasets, we build a synthetic dataset N-FineGym and a real dataset E-FAction. The N-FineGym dataset is an event-based extension of FineGym [12]. Videos from FineGym are converted to event stream with ESIM [41] and SuperSloMo [42] (refer to our project page). We follow the RGB frame few-shot works [15] to use a split of 72, 13 and 14 classes in the training, validation and testing sets. For our E-FAction, we split coarse actions into 13, 1, and 1 classes for training, validation and testing sets, and use the fine-grained classes from the coarse classes for few-shot FGH action recognition. We further compare our approach on published action recognition datasets. The DailyAction [34] dataset is captured with the DAVIS128 event camera. We randomly split 7 and 5 classes for training and testing. There are only 12 classes, and we do not use a validation split. The $\text{THU}^{\text{E-ACT}}_{-50}$ [16] dataset captured with the CeleX-V camera is split into 35 training, 5 validation, and 10 testing classes. When training our event feature extractor, we use the event and paired RGB frames in the training set.

B. Implementation Details

Architectural Detail. We use a pre-trained ViT-B/16 [43] from CLIP [13] for $\mathbf{E}^{\text{clip}}(\cdot)$, and layers before the FlowHead

from pre-trained Videoflow [14] for $\mathbf{E}^{\text{vif}}(\cdot)$. The $\mathbf{H}^{\text{clip}}(\cdot)$ and $\mathbf{H}^{\text{vif}}(\cdot)$ are each implemented by a feedforward network with an average pooling layer. For $\mathbf{E}^{\text{evt}}(\cdot)$, we use an 18-layer Resnet3D [44]. For few-shot FGH action recognition, we follow architectures of state-of-the-art methods, and replace the default CLIP image encoder with our learned $\mathbf{E}^{\text{evt}}(\cdot)$.

Training Setup. To train our $\mathbf{E}^{\text{evt}}(\cdot)$, we use an input resolution of 224×224, with $T = 8$ and $\lambda = 0.1$ for 300 epochs using SGD optimizer. The learning rate, momentum, and weight decay are 0.01, 0.9, and 0.01. Then we train for an additional 100 epochs using AdamW optimizer with a learning rate of 0.001 and a weight decay of 0.01. All experiments in this study are conducted on a single NVIDIA RTX 3090.

TABLE II

PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ACROSS THE FOUR EVENT-BASED HUMAN ACTION RECOGNITION DATASETS. THE BEST PERFORMANCE IS IN BOLD.

Model	N-FineGym	E-FAction	DailyAction	$\text{THU}^{\text{E-ACT}}_{-50}$
CLIP-L [13]	60.1	55.8	89.2	71.3
ResNet3D-N [45]	67.1	58.1	93.9	76.4
ResNet3D-K [46]	70.2	59.6	89.3	76.0
MASTAF [10]	63.4	60.1	95.1	82.3
Shi *et al.* [15]	70.6	60.5	93.5	79.2
Ours	**73.3**	**62.0**	**97.6**	**84.8**

TABLE III

PERFORMANCE COMPARISON OF USING OUR AND THE DEFAULT FEATURE EXTRACTORS (FE) IN STATE-OF-THE-ART FEW-SHOT METHODS.

FE	Method	N-FineGym	E-FAction	DailyAction	$\text{THU}^{\text{E-ACT}}_{-50}$
Default	MASTAF	63.4	60.1	95.1	82.3
Ours	MASTAF	65.8	66.7	96.4	88.4
Default	Shi *et al.*	70.6	60.5	93.5	79.2
Ours	Shi *et al.*	73.3	62.0	97.6	84.8

C. Results

State-of-the-art Methods. We compare with the state-of-the-art few-shot action recognition methods: MASTAF [10], and Shi *et al.* [15]. MASTAF is a model-agnostic network using attention for video classification, with a fusion network and nearest neighbor classifier. Shi *et al.* use a vision-language model and lightweight temporal network to classify actions from text-based knowledge.

Baselines. We use the image encoder from CLIP [13] with logistic regression for few-shot action recognition. We replace the CLIP image encoder from the state-of-the-art method (Shi *et al.* [15]) with the N-ImageNet [45] and Kinetic-400 [46] dataset pre-trained ResNet3D [44]. We call them CLIP-L, ResNet3D-N, and ResNet3D-K.

We compare with state-of-the-art and baseline methods for 5-way 5-shot FGH action recognition in Tab. II. Please refer to [10], [15] for more details. Our method has the best performance across the four datasets. Our findings are described in the following. i) Compared to Shi *et al.* [15], we share the few-shot recognition method but a different feature extractor, and our event feature extractor has 2.7%, 1.5%, 4.1%, and 5.6% higher accuracy than the default feature extractor of Shi *et al.* [15]. Our performance shows that there is a significant domain gap between the event and RGB domain, indicating that the event feature extractor is crucial for few-shot FGH action recognition, especially for challenging scenarios. ii) For ResNet3D-N and ResNet3D-K, the semantic classes of event stream are used to train the event feature extractor. Using the same few-shot method and feature extractor architectures, ResNet3D-N and ResNet3D-K have a lower accuracy than ours, because our method using pre-trained CLIP [13] and VideoFlow [14] for supervision is more robust and effective than supervision with class labels in learning a feature extractor for event stream. iii) Our method has a better performance than CLIP-L and Shi *et al.* [15]. Our event feature extractor is learned from the image encoder of CLIP, while CLIP-L and Shi *et al.* [15] directly use the image encoder of CLIP. Using RGB frames as an intermediate step to learn a feature extractor for event stream can help transfer knowledge for FGH action recognition.

In Tab. III, we compare our and the default feature extractors (CLIP [13] and ResNet3D) in state-of-the-art few-shot FGH action recognition methods [10], [15]. With our event feature extractor, the few-shot recognition performance is consistently improved, which shows that our extracted features can be extensively used for FGH action recognition methods.

TABLE IV

ABLATION OF LOSS FUNCTIONS ON LEARNING OUR EVENT FEATURE EXTRACTOR.

$\mathcal{L}_{\text{m}}$	$\mathcal{L}_{\text{s}}$	$\mathcal{L}^{\text{cor}}_{\text{m}}$	$\mathcal{L}^{\text{kl}}_{\text{s}}$	N-FineGym	E-FAction	DailyAction	$\text{THU}^{\text{E-ACT}}_{-50}$
✓	✗	✗	✗	70.3	60.1	95.2	82.0
✗	✓	✗	✗	72.1	61.5	96.4	83.5
✓	✓	✗	✗	**73.3**	**62.0**	**97.6**	**84.8**
✓	✗	✗	✓	71.1	61.2	92.5	76.6
✗	✓	✓	✗	72.3	61.4	95.8	81.8
✗	✗	✓	✗	64.3	57.1	86.5	75.2
✗	✗	✗	✓	69.2	60.2	87.1	75.8
✗	✗	✓	✓	69.9	60.8	90.4	76.3

D. Ablation Studies

We study the losses for learning our event feature extractor. In Tab. IV, using only the $\mathcal{L}_{\text{m}}$/$\mathcal{L}_{\text{s}}$ have 70.3%/72.1%, 60.1%/61.5%, 95.2%/96.4%, and 82.0%/83.5% accuracy on the N-FineGym [12], E-FAction, DailyAction [34], and $\text{THU}^{\text{E-ACT}}_{-50}$ [16] datasets. When combining $\mathcal{L}_{\text{m}}$ and $\mathcal{L}_{\text{s}}$, we consistently improve the performance of only using $\mathcal{L}_{\text{m}}$ /$\mathcal{L}_{\text{s}}$ by 3.0%/1.2%, 1.9%/0.5%, 2.4%/1.2%, and 2.8%/1.3%.

We then study two variations of $\mathcal{L}_{\text{m}}$ and $\mathcal{L}_{\text{s}}$ about correlations. i) $\mathcal{L}^{\text{cor}}_{\text{m}}$ calculates the KL loss on the motion features, with using a temporal correlation. ii) $\mathcal{L}^{\text{kl}}_{\text{s}}$ calculates the KL loss directly at the temporal dimension of semantic features, without using the the self-correlation. We have the best

TABLE V

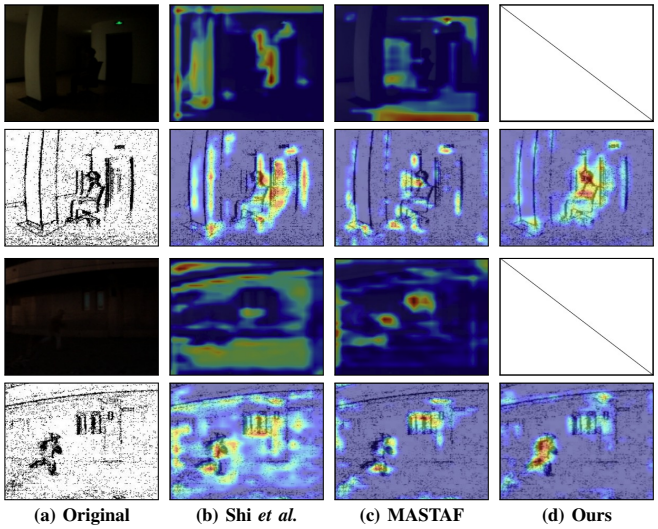
PERFORMANCE COMPARISON BETWEEN RGB AND EVENT FRAMES WITH LOW LIGHTING CONDITIONS.

Method	RGB	Event
Shi *et al.*	56.1	61.5
MASTAF	60.5	64.5
Ours	/	**66.0**

model by $\mathcal{L}_{\mathrm{m}}$ and $\mathcal{L}_{\mathrm{s}}$ with higher accuracy than using $\mathcal{L}_{\mathrm{m}}^{\mathrm{cor}}$ and $\mathcal{L}_{\mathrm{s}}^{\mathrm{kl}}$. Motion features are already temporally extracted by the pre-trained VideoFlow, and relative to temporal nearby frames. A temporal self-correlation in $\mathcal{L}_{\mathrm{m}}^{\mathrm{cor}}$ changes the motion features to features related to relative speed, and is less useful than motion features for learning an event feature extractor for FGH action recognition. For semantic features, they are independently extracted by the pre-trained CLIP image encoder at the temporal dimension. Fine-grained action recognition is about learning semantic variations along the temporal dimension, which can be addressed by using a temporal self-correlation on the semantic features in $\mathcal{L}_{\mathrm{s}}$.

E. Performance on Low Lighting Conditions

To explore the high dynamic range of event cameras in FGH action recognition, we use event and RGB frames with low light conditions for state-of-the-art methods. We compare the performance of the state-of-the-art approaches with ours on the sub-set of our E-FAction dataset where data is captured under low lighting conditions. In Tab. V, all methods have best accuracy with event frames. The performance underscores event stream importance and potential in action recognition with challenging lighting conditions. To further illustrate the advantage, we provide heat maps for event and RGB frames, as shown in Fig. 5. The heat maps indicate that it is hard for models to recognize actions with features extracted from RGB frames.



(a) Original **(b) Shi *et al.*** **(c) MASTAF** **(d) Ours**

Fig. 5. Heat maps for models trained with RGB and event frames. (a) are paired RGB and event frames, and (b)-(d) show heat maps of Shi *et al.*, MASTAF, and our method. Our event feature extractor makes it easier for the model to focus on human body movement areas. The first and third rows of (d) do not have responses because our method works only for events. Best viewed in colour on the screen.

VI. CONCLUSIONS

In this paper, we introduce the first event dataset for fine-grained action recognition. It has 3304 paired 'event stream and RGB sequence' and 128 fine-grained action classes. For human-robot interaction within open-set real-world environments, we learn a feature extractor for event frames in few-shot FGH action recognition. We use pre-trained VideoFlow encoder and CLIP image encoder to extract motion and semantic features of RGB frames, and design a temporal motion and a temporal semantic loss to supervise the feature extractor for event frames. On synthetic and real event datasets, our event feature extractor consistently achieves state-of-the-art few-shot FGH action recognition performance. In future work, we plan to use higher resolution event cameras to capture finer motion details for improved evaluations.

ACKNOWLEDGMENT

This work was supported in part by the Beijing Institute of Technology Research Fund Program for Young Scholars, BIT Special-Zone, and National Natural Science Foundation of China 62302045.

REFERENCES

[1] T. de Blegiers, I. R. Dave, A. Yousaf, and M. Shah, "Eventtransact: A video transformer-based framework for event-camera based action recognition," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*. IEEE, 2023, pp. 1–7.

[2] X. Wang, S. Zhang, Z. Qing, C. Gao, Y. Zhang, D. Zhao, and N. Sang, "Molo: Motion-augmented long-short contrastive learning for few-shot action recognition," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 18 011–18 021.

[3] Y. Li, Y. Li, and N. Vasconcelos, "Resound: Towards action recognition without representation bias," in *Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 513–528.

[4] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, *et al.*, "Event-based vision: A survey," *pami*, vol. 44, no. 1, pp. 154–180, 2020.

[5] G. Diraco, G. Rescio, P. Siciliano, and A. Leone, "Review on human action recognition in smart living: Sensing technology, multimodality, real-time processing, interoperability, and resource-constrained processing," *Sensors*, vol. 23, no. 11, p. 5281, 2023.

[6] S. Ahmad, G. Scarpellini, P. Morerio, and A. Del Bue, "Event-driven re-id: A new benchmark and method towards privacy-preserving person re-identification," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 459–468.

[7] G. Lenz, S.-H. Ieng, and R. Benosman, "Event-based face detection and tracking using the dynamics of eye blinks," *Frontiers in Neuroscience*, vol. 14, p. 587, 2020.

[8] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 × 128 120 db 15 us latency asynchronous temporal contrast vision sensor," *Solid-State Circuits, IEEE Journal of*, vol. 43, pp. 566 – 576, 03 2008.

[9] D. Gehrig, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Video to events: Recycling video datasets for event cameras," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 3586–3595.

[10] X. Liu, H. Zhang, H. Pirsiavash, and X. Liu, "Mastaf: A model-agnostic spatio-temporal attention fusion network for few-shot video classification," in *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2023, pp. 2507–2516.

[11] Y. Yang, L. Pan, and L. Liu, "Event camera data pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 699–10 709.

[12] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 2616–2625.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learning (ICML)*. PMLR, 2021, pp. 8748–8763.

[14] X. Shi, Z. Huang, W. Bian, D. Li, M. Zhang, K. C. Cheung, S. See, H. Qin, J. Dai, and H. Li, "Videoflow: Exploiting temporal cues for multi-frame optical flow estimation," *arXiv preprint arXiv:2303.08340*, 2023.

[15] Y. Shi, X. Wu, and H. Lin, "Knowledge prompting for few-shot action recognition," *arXiv preprint arXiv:2211.12030*, 2022.

[16] Y. Gao, J. Lu, S. Li, N. Ma, S. Du, Y. Li, and Q. Dai, "Action recognition and benchmark using event cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14 081–14 097, 2023.

[17] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman, "Hots: a hierarchy of event-based time-surfaces for pattern recognition," *pami*, vol. 39, no. 7, pp. 1346–1359, 2016.

[18] C. Huang, "Event-based action recognition using timestamp image encoding network," *arXiv preprint arXiv:2009.13049*, 2020.

[19] X. Wu and J. Yuan, "Multipath event-based network for low-power human action recognition," in *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*. IEEE, 2020, pp. 1–5.

[20] C. Plizzari, M. Planamente, G. Goletto, M. Cannici, E. Gusso, M. Matteucci, and B. Caputo, "E2 (go) motion: Motion augmented event stream for egocentric action recognition," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 19 935–19 947.

[21] J. Chen, J. Meng, X. Wang, and J. Yuan, "Dynamic graph cnn for event-camera based gesture recognition," in *IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2020, pp. 1–5.

[22] Y. Gao, J. Lu, S. Li, N. Ma, S. Du, Y. Li, and Q. Dai, "Action recognition and benchmark using event cameras," *pami*, 2023.

[23] X. Wang, Z. Wu, B. Jiang, Z. Bao, L. Zhu, G. Li, Y. Wang, and Y. Tian, "Hardvs: Revisiting human activity recognition with dynamic vision sensors," *arXiv preprint arXiv:2211.09648*, 2022.

[24] E. Ceolini, C. Frenkel, S. B. Shrestha, G. Taverni, L. Khacef, M. Payvand, and E. Donati, "Hand-gesture recognition based on emg and event-based camera sensor fusion: A benchmark in neuromorphic computing," *Frontiers in neuroscience*, vol. 14, p. 520438, 2020.

[25] Q. Liu, G. Pan, H. Ruan, D. Xing, Q. Xu, and H. Tang, "Unsupervised aer object recognition based on multiscale spatio-temporal features and spiking neurons," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5300–5311, 2020.

[26] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[27] L. Zhu and Y. Yang, "Label independent memory for semi-supervised few-shot video classification," *pami*, vol. 44, no. 1, pp. 273–285, 2020.

[28] K. Cao, J. Ji, Z. Cao, C.-Y. Chang, and J. C. Niebles, "Few-shot video classification via temporal alignment," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 10 618–10 627.

[29] Z. Zhu, L. Wang, S. Guo, and G. Wu, "A closer look at few-shot video classification: A new baseline and benchmark," *arXiv preprint arXiv:2110.12358*, 2021.

[30] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, *et al.*, "A low power, fully event-based gesture recognition system," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 7243–7252.

[31] Y. Bi, A. Chadha, A. Abbas, E. Bourtsoulatze, and Y. Andreopoulos, "Graph-based spatio-temporal feature learning for neuromorphic vision sensing," *IEEE Trans. Image Process.*, vol. 29, pp. 9084–9098, 2020.

[32] S. Miao, G. Chen, X. Ning, Y. Zi, K. Ren, Z. Bing, and A. Knoll, "Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection," *Frontiers in neurorobotics*, vol. 13, p. 38, 2019.

[33] E. Calabrese, G. Taverni, C. Awai Easthope, S. Skriabine, F. Corradi, L. Longinotti, K. Eng, and T. Delbruck, "Dhp19: Dynamic vision sensor 3d human pose dataset," in *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, 2019, pp. 0–0.

[34] Q. Liu, D. Xing, H. Tang, D. Ma, and G. Pan, "Event-based action recognition using motion information and spiking neural networks." in *IJCAI*, 2021, pp. 1743–1749.

[35] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *Int. Conf. Comput. Vis. (ICCV)*. IEEE, 2011, pp. 2556–2563.

[36] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2014, pp. 740–755.

[38] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[39] Z. Wang, F. C. Ojeda, A. Bisulco, D. Lee, C. J. Taylor, K. Daniilidis, M. A. Hsieh, D. D. Lee, and V. Isler, "Ev-catcher: High-speed object catching using low-latency event-based neural networks," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 8737–8744, 2022.

[40] Y. Tang, B. Béjar, and R. Vidal, "Semantic-aware video representation for few-shot action recognition," in *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2024, pp. 6458–6468.

[41] H. Rebecq, D. Gehrig, and D. Scaramuzza, "Esim: an open event camera simulator," in *Conference on robot learning*. PMLR, 2018, pp. 969–982.

[42] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super slomo: High quality estimation of multiple intermediate frames for video interpolation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 9000–9008.

[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[44] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 6450–6459.

[45] J. Kim, J. Bae, G. Park, D. Zhang, and Y. M. Kim, "N-imagenet: Towards robust, fine-grained object recognition with event cameras," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 2146–2156.

[46] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.