

## EZSR: Event-based Zero-Shot Recognition

Yan Yang<sup>†</sup> Liyuan Pan<sup>†</sup>✉ Dongxu Li<sup>‡</sup> Liu Liu<sup>§</sup>  
<sup>†</sup>BITSZ & School of CSAT, BIT <sup>‡</sup>ANU <sup>§</sup>KooMap, Huawei  
 Project Page: <https://yan98.github.io/EZSR/>

Yan.Yang@anu.edu.au Liyuan.Pan@bit.edu.cn dongxuli1005@gmail.com liuliu33@huawei.com

### Abstract

This paper studies zero-shot object recognition using event camera data. Guided by CLIP, which is pre-trained on RGB images, existing approaches achieve zero-shot object recognition by optimizing embedding similarities between event data and RGB images respectively encoded by an event encoder and the CLIP image encoder. Alternatively, several methods learn RGB frame reconstructions from event data for the CLIP image encoder. However, they often result in suboptimal zero-shot performance.

This study develops an event encoder without relying on additional reconstruction networks. We theoretically analyze the performance bottlenecks of previous approaches: the embedding optimization objectives are prone to suffer from the spatial sparsity of event data, causing semantic misalignments between the learned event embedding space and the CLIP text embedding space. To mitigate the issue, we explore a scalar-wise modulation strategy. Furthermore, to scale up the number of events and RGB data pairs for training, we also study a pipeline for synthesizing event data from static RGB images in mass.

Experimentally, we demonstrate an attractive scaling property in the number of parameters and synthesized data. We achieve superior zero-shot object recognition performance on extensive standard benchmark datasets, even compared with past supervised learning approaches. For example, our model with a ViT/B-16 backbone achieves 47.84% zero-shot accuracy on the N-ImageNet dataset.

### 1. Introduction

An event camera asynchronously captures pixel-wise brightness changes as an event stream, each recording the position, time, and polarity of a brightness change. Compared to traditional image sensors, event cameras not only offer benefits such as motion blur-free imaging and high

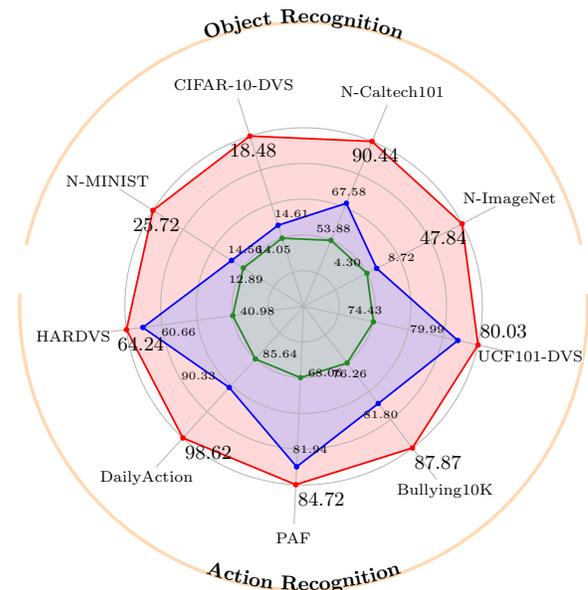


Figure 1. Comparison of **our** accuracies (%) with respect to the **second-best** and **third-best** accuracies (%) from previous methods [44, 50, 51] on object [20, 23, 32, 33] and action [4, 11, 27, 30, 43] recognition. Beside each axis, the dataset name is given.

temporal resolution but also consume less energy and are robust under adverse lighting conditions. The significant advantages of event cameras have led to extensive neural network applications, such as object recognition [20, 51], semantic segmentation [22, 47], etc. [46].

Aiming a generalized object recognition in open-world settings, this paper studies zero-shot object recognition using event camera data; in other words, our method can be tested on object classes that are not seen in the training set. Our method is contrastive learning-based, and benefited from distillation of pre-trained zero-shot object recognition frameworks for RGB images. We particularly focus on CLIP [35, 40], because the extraordinary success of CLIP has been demonstrated in extensive tasks.

To adapt the existing zero-shot RGB network, CLIP [50],

✉ Corresponding author.

to the event domain, a direct approach is to convert the event data into an event frame, and treat it as an RGB image for zero-shot object recognition [44, 50]. However, event frames are typically spatially sparse, whereas RGB images densely record light intensity of a scene. The domain divergence undermines the performance of CLIP. In extensions, grayscale images are reconstructed from event data by learning an additional network [8], enabling compatibility between the CLIP image encoder and the event data. However, it incurs extra computational costs during inference, and usually results in poor performance due to the low reconstruction quality and error accumulation between the reconstruction and recognition stages.

Therefore, some works distill an event encoder from the CLIP image encoder by using paired event data and RGB images [51]. Afterwards, object recognition is computed from the event encoder and the CLIP text encoder. Contrastive learning is usually employed for learning the event encoder. However, it not only optimizes similarities between paired event and RGB data but also dissimilarities between non-paired ones. Due to the spatial sparsity of event data, event embeddings tend to become overly similar (Fig. 2). When contrasting with an image, this leads to similar similarities between the image and both paired and non-paired events, underscoring optimization attention to ensure the embedding discriminativeness. Theoretically, it results in the degree of freedom for alignments between event embeddings and text embeddings from the CLIP text encoder (Sec. 3). This misalignment hampers effective event-based zero-shot learning. Some works use text embeddings of class names into training, yet show little zero-shot performance. Moreover, due to the lack of paired event and RGB datasets, the previous approach usually performs training and testing on the same dataset [8, 51], leading an biased evaluation.

This paper designs an event-based zero-shot object recognition framework for mitigating the above drawbacks. Additionally, our method demonstrates the appealing scalability in terms of model parameters and training data. The favorable properties are credited to two key designs: scalar-wise modulation and data synthesis.

We theoretically and experimentally show that objectives involving dissimilarities optimization between non-paired event and RGB embedding lead to the semantic misalignment. Then, we propose a scalar-wise modulation strategy to directly align the event to RGB data embeddings, rather than solely relying on the contrastive learning objective. It compels the network to adaptively mine scalar-wise semantic alignments between RGB and event embeddings, dynamically directing optimization attention toward distinguishing overly similar event embeddings. These scalar-wise alignments transfer RGB and text data alignments directly to the event and text data, and prompt the zero-shot

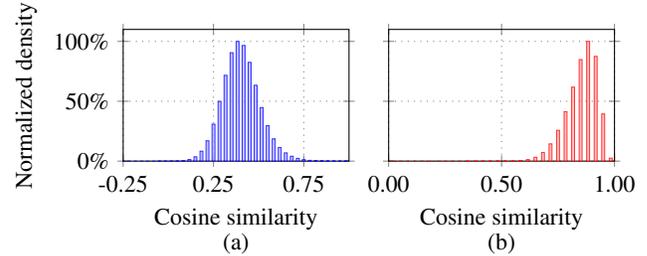


Figure 2. Similarity distribution for (a) RGB and (b) event embeddings. RGB and event embeddings are extracted using a pre-trained CLIP model [13] on the validation fold of ImageNet-1K [10] and N-ImageNet [20] datasets. Cosine similarities are computed separately among the RGB and event embeddings, and the density is normalized by the maximum value.

object recognition with the CLIP text encoder.

To overcome the dataset scarcity challenge, we use synthetic data for training our network. Our target is zero-shot object recognition that typically relies on event data in a short duration (*e.g.*, dozens of milliseconds) and the event data in the duration usually contains linear motion [25]. We randomly generate affine transformations and interpolate static RGB images into a sequence for event data synthesis. Compared to traditional methods that rely on video and pre-trained frame interpolation networks, our approach shows a low computation complexity, and introduces greater diversity than the same amount of video data. Our contributions are summarized as follows:

- A scalable zero-shot object recognition framework for event camera data;
  - A scalar-wise modulation strategy to promote alignments between embeddings from the event and text encoders;
  - A large-scale and diverse dataset of event and RGB pairs.
- Experimentally, we evaluate our network on nine standard event datasets, demonstrating competitive performance even compared to the dataset-specific methods. Refer to Fig. 1 for an overall comparison. Our code and dataset will be made publicly available.

## 2. Related Works

We review the recent achievements of event camera data in zero-shot learning and relevant self-supervised learning methods. Then, we introduce the event datasets.

**Zero-shot learning.** Event-based zero-shot learning is still in an early stage, with studies focused on object recognition [8, 44, 50, 51] and semantic segmentation [22] using CLIP [35]. We are interested in zero-shot object recognition that can be broadly categorized into two groups: i) reconstruction-based approach. With extra computational costs, it reconstructs grayscale images from event data to leverage pre-trained CLIP for object recognition. By exploiting the strong RGB domain prior from CLIP, the recon-

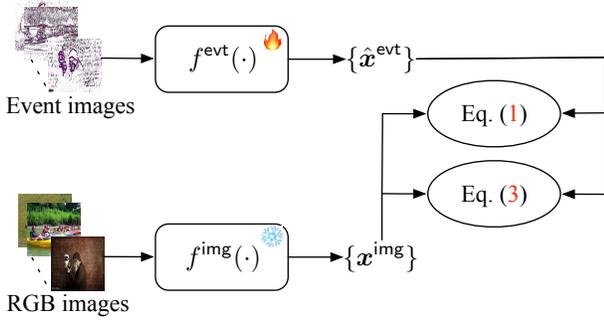


Figure 3. Overview of our method. Our goal is to learn an event encoder  $f^{evt}(\cdot)$  to replace the image encoder  $f^{img}(\cdot)$  from a pre-trained CLIP for allowing zero-shot object recognition with event data. Given paired event and RGB images, we respectively extract the embeddings  $\{\hat{x}^{evt}\}$  and  $\{x^{img}\}$  from  $f^{evt}(\cdot)$  and  $f^{img}(\cdot)$  to optimize Eq. (1) and Eq. (3). The fire (i.e., 🔥) and snowflake (i.e., ❄️) emojis respectively denote trainable and frozen components.

struction network can be trained by using only event data [8]. Although it offers promising reconstruction quality, the zero-shot recognition performance is poor due to error accumulation; ii) contrastive learning-based approach. Studies like [50, 51] enforce embedding similarities between an event encoder and the pre-trained image encoder of CLIP, using paired event data and RGB images for training. The primary objective is instance discrimination between event and RGB images through contrastive learning. However, as analyzed in Sec. 1, their performance is hindered, and often requires sacrificing large-scale datasets intended for evaluation to be used for training. Other methods [44, 51] convert event data into event frames to explore the direct use of CLIP for zero-shot recognition or few-shot learning. This paper studies a method trained on synthetic data to address their limitations and improve zero-shot object recognition.

**Self-supervised learning.** Similar to zero-shot learning, self-supervised learning aims to learn generic feature embeddings of event data. It employs tasks like masked image modeling [17, 21], contrastive learning [7, 47], and self-distillation [31, 46]. These methods improve performance on downstream tasks through transfer learning. However, unlike zero-shot learning, they are required to learn a decoding head for recognition. This paper focuses on zero-shot object recognition on event data, enabling recognition with a simple forward pass without further training.

**Event datasets.** Event cameras offer high temporal resolution, reduced motion blur, and lower power consumption, making them ideal for dynamic and low-light environments. This has spurred the development of various event-based datasets for object and action recognition. Notable object recognition datasets include N-ImageNet [20], N-Cars [39], CIFAR10-DVS [23], N-Caltech101 [32], DVS-128-Gesture [1], etc. [3, 33, 36]. Examples of the action

recognition datasets are HMDB-DVS [4], UCF-DVS [4], PAF [30], DailyAction [27], and HARDVS [43]. This paper extensively evaluates our method on object recognition and action recognition datasets.

### 3. Method

We first introduce the background of contrastive learning by establishing a baseline. Next, our framework is presented component by component, tracking the performance variations of the baseline. All training is conducted on our synthetic dataset, and the N-ImageNet, the largest event-based recognition dataset, is used for benchmarking.

#### 3.1. Background

**Preliminary.** Contrastive learning pulls embeddings of the same instance close to each other and pushes apart embeddings of different instances. It embeds instances into a query set  $\{q\}$  and a key set  $\{k\}$ . For each query embedding  $q$ , there is a matching key embedding  $k_+$  and a set of non-matching key embeddings  $\{k_-\}$ . Then, an embedding space is optimized to ensure that  $q$  is close to  $k_+$  and distant from  $\{k_-\}$  through minimizing an InfoNCE loss [42]. The InfoNCE loss for a query embedding  $q$  is given by

$$\mathcal{L}_{nce}(q, \{k\}) = -\log \frac{\exp(q \cdot k_+ / \tau)}{\exp(q \cdot k_+ / \tau) + \sum_{k_-} \exp(q \cdot k_- / \tau)},$$

where  $\tau$  is a temperature parameter to control the distribution sharpness. Usually, the query  $q$  and key  $k$  embeddings are  $\ell_2$ -normalized, and the dot product  $\cdot$  between them calculates their cosine similarity.

#### 3.2. Baseline

In the context of learning an event encoder using a pre-trained CLIP image encoder, paired event data and RGB images are treated as the same instances. The collections of event and RGB data are respectively encoded into two embedding sets:  $\{\hat{x}^{evt}\}$  for events and  $\{x^{img}\}$  for images, with an event encoder  $f^{evt}(\cdot)$  and a pre-trained CLIP image encoder  $f^{img}(\cdot)$ . The embeddings  $\{\hat{x}^{evt}\}$  and  $\{x^{img}\}$  are mutually served as query set and key sets for optimizing a symmetrized InfoNCE loss,

$$\mathcal{L}_{baseline} = \frac{1}{|\{\hat{x}^{evt}\}|} \sum_{\hat{x}^{evt} \in \{\hat{x}^{evt}\}} \mathcal{L}_{nce}(\hat{x}^{evt}, \{x^{img}\}) + \frac{1}{|\{x^{img}\}|} \sum_{x^{img} \in \{x^{img}\}} \mathcal{L}_{nce}(x^{img}, \{\hat{x}^{evt}\}), \quad (1)$$

where  $|\{\hat{x}^{evt}\}|$  and  $|\{x^{img}\}|$  respectively denote the sizes of event and image embedding sets. During training, the image encoder  $f^{img}(\cdot)$  is frozen, maintaining the CLIP embedding

Table 1. Zero-shot object recognition accuracy (%) on the N-ImageNet dataset [20] by combining our model components.

$\mathcal{L}_{\text{baseline}}$	Remark 1	Scalar-wise Modulation	Accuracy
✓	✗	✗	9.57
✓	✓	✗	43.48
✓	✗	✓	47.84
✗	✗	✓	47.80
✗	✓	✓	48.63
✓	✓	✓	48.86

space structure for directly using the pre-trained CLIP text encoder  $f^{\text{txt}}(\cdot)$  in zero-shot object recognition.

Though the baseline approach can effectively minimize Eq. (1) during training, it fails to find an accurate zero-shot object recognition performance, as seen in Tab. 1. The low performance, 9.57%, suggests a divergence between the optimization and evaluation objectives.

We note the divergence is raised by the degree of freedom associated with the embedding dimension and optimization objective. Let  $\mathbf{x}_+^{\text{img}}$  and  $\mathbf{x}_+^{\text{txt}}$  embedded respectively by the pre-trained  $f^{\text{img}}(\cdot)$  and  $f^{\text{txt}}(\cdot)$  are matching RGB image and text descriptions, which are also matching with  $\hat{\mathbf{x}}^{\text{evt}}$ , and  $\hat{\mathbf{x}}_-^{\text{evt}}$  and  $\mathbf{x}_-^{\text{img}}$  be the non-matching event and image embedding.

**Lemma 1.** When Eq. (1) is effectively minimized,  $\mathbf{x}_+^{\text{img}} \cdot \mathbf{x}_+^{\text{txt}} > \mathbf{x}_-^{\text{img}} \cdot \mathbf{x}_+^{\text{txt}}$  does not imply  $\hat{\mathbf{x}}^{\text{evt}} \cdot \mathbf{x}_+^{\text{txt}} > \hat{\mathbf{x}}_-^{\text{evt}} \cdot \mathbf{x}_+^{\text{txt}}$ , due to degree of freedom in the embedding space.

**Remark 1.** The misalignment between event and text embeddings can be mitigated if a proper reference dataset is provided. According to Eq. (1), we know the event embedding is pulled to semantic similar image embeddings, while the image embedding is already aligned with the text embedding in the CLIP for zero-shot object recognition. Thus, we translate  $\hat{\mathbf{x}}^{\text{evt}}$  by using a pool of pre-embedded image embeddings  $\{\mathbf{x}^{\text{img}}\}$  to align with the CLIP text embedding. Let  $\{k\}$  be the indices set of  $k$ -NN embedding from  $\{\mathbf{x}^{\text{img}}\}$ , using cosine similarities as the distance measure. The translated event embedding  $\tilde{\mathbf{x}}^{\text{evt}}$  is

$$\tilde{\mathbf{x}}^{\text{evt}} = \sum_{k \in \{k\}} \frac{\hat{\mathbf{x}}^{\text{evt}} \cdot \mathbf{x}_k^{\text{img}} + 1}{\sum_{k' \in \{k\}} (\hat{\mathbf{x}}^{\text{evt}} \cdot \mathbf{x}_{k'}^{\text{img}} + 1)} \mathbf{x}_k^{\text{img}}. \quad (2)$$

With the translated event embedding  $\tilde{\mathbf{x}}^{\text{evt}}$ , the zero-shot object recognition performance can be improved from 9.57% to 43.48%. However, we do not focus on the strategies, as the performance is usually reference dataset dependent. For readers' interest, we present the ablations with it in Tab. 1.

### 3.3. Scalar-wise Modulation

One may note that there is an exemption in Lem. 1. If  $\hat{\mathbf{x}}^{\text{evt}} \cdot \mathbf{x}_+^{\text{img}} = 1$ , this would imply perfect alignment between  $\hat{\mathbf{x}}^{\text{evt}}$

and  $\mathbf{x}_+^{\text{img}}$ . In this case,  $\hat{\mathbf{x}}^{\text{evt}}$  would inherently satisfy  $\hat{\mathbf{x}}^{\text{evt}} \cdot \mathbf{x}_+^{\text{txt}} > \hat{\mathbf{x}}_-^{\text{evt}} \cdot \mathbf{x}_+^{\text{txt}}$  if  $\mathbf{x}_+^{\text{img}} \cdot \mathbf{x}_+^{\text{txt}} > \mathbf{x}_-^{\text{img}} \cdot \mathbf{x}_+^{\text{txt}}$ .

Inspired by the observation, rather than simply enforcing the embedding similarities between  $\hat{\mathbf{x}}^{\text{evt}}$  and  $\mathbf{x}_+^{\text{img}}$ , we introduce a scalar-wise modulation to enforce the scalar-wise alignments between  $\hat{\mathbf{x}}^{\text{evt}}$  and  $\mathbf{x}_+^{\text{img}}$ . As noted in Sec. 1, event embeddings  $\{\hat{\mathbf{x}}^{\text{evt}}\}$  tend to be overly similar to one another, warranting greater optimization focus. To address it, our approach is adaptive, constraining the degrees of freedom by ensuring that each scalar of the event embedding and the image embedding are aligned, therefore enhancing the embedding alignment.

We achieve it by adding a mean squared error to the optimization objective that adaptively penalizes the scalar-wise deviations in the embeddings. The loss is given by

$$\mathcal{L}_{\text{mod}} = \sum_{\lambda_{\hat{\mathbf{x}}^{\text{evt}}}, \hat{\mathbf{x}}^{\text{evt}}, \mathbf{x}_+^{\text{img}}} \lambda_{\hat{\mathbf{x}}^{\text{evt}}} \|\hat{\mathbf{x}}^{\text{evt}} - \mathbf{x}_+^{\text{img}}\|^2, \quad (3)$$

$$\lambda_{\hat{\mathbf{x}}^{\text{evt}}} = \mathcal{N}\left(1 - \frac{\lambda_{\hat{\mathbf{x}}^{\text{evt}}}^{\text{unf}} - \min(\{\lambda_{\hat{\mathbf{x}}^{\text{evt}}}^{\text{unf}}\})}{\max(\{\lambda_{\hat{\mathbf{x}}^{\text{evt}}}^{\text{unf}}\}) - \min(\{\lambda_{\hat{\mathbf{x}}^{\text{evt}}}^{\text{unf}}\})}\right), \quad (4)$$

$$\lambda_{\hat{\mathbf{x}}^{\text{evt}}}^{\text{unf}} = \frac{\sum_{\hat{\mathbf{x}}_-^{\text{evt}}} (\hat{\mathbf{x}}^{\text{evt}} \cdot \hat{\mathbf{x}}_-^{\text{evt}} + 1)}{\sum_{\hat{\mathbf{x}}_-^{\text{evt}'}} \sum_{\hat{\mathbf{x}}_-^{\text{evt}'}} (\hat{\mathbf{x}}^{\text{evt}'} \cdot \hat{\mathbf{x}}_-^{\text{evt}'} + 1)}, \quad (5)$$

where  $\mathcal{N}(\cdot)$  is a Gaussian distribution function, and the subscript ‘-’ consistently to denote negative matching embedding (i.e., another embedding in the batch  $\{\hat{\mathbf{x}}^{\text{evt}}\}$ ).  $\|\cdot\|$  is the Frobenius norm that reduces the vector to a scalar. The input to the Gaussian function falls within the range from 0 to 1, ensuring that  $\lambda_{\hat{\mathbf{x}}^{\text{evt}}}$  remains non-negative. By framing  $\mathcal{L}_{\text{baseline}}$  as a regularization term to maintain embedding discriminativeness, the overall objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{baseline}} + \mathcal{L}_{\text{mod}}. \quad (6)$$

With the modulation, the degrees of freedom are constrained, taming the event embeddings to effectively align with the text embeddings for improving the zero-shot performance.

We present the detailed combinations of using  $\mathcal{L}_{\text{mod}}$  in Tab. 1. It consistently improves the performance by enforcing a stronger alignment constraint. For example, 47.84% accuracy is achieved with our scalar-wise modulation.

### 3.4. Data Synthesis

Due to the lack of paired event data and RGB images for training, we generate synthetic event data. In object recognition tasks using event datasets, events captured over a short duration are typically used, emphasizing semantic information rather than motion information. Since events within a short duration often exhibit linear motion [25], we simulate random linear motion on a static RGB image. The RGB image is warped to create a video, and the video is

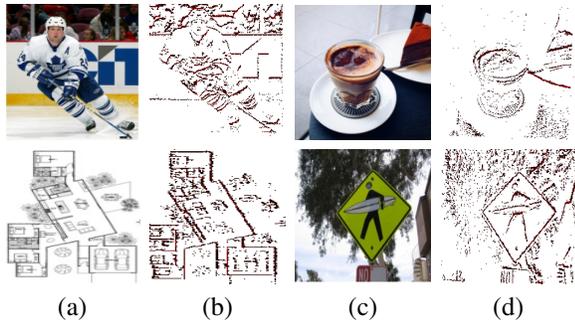


Figure 4. Samples of synthetic event data. (a)/(c) are RGB images. (b)/(d) are event frames, where red and blue indicate positive and negative events, respectively.

subsequently transformed into event data using V2E [18]. Three motion patterns are explored, as described below:

- Translation. Random horizontal and vertical maximum displacements are first sampled, and then the static image is displaced linearly to meet the maximum displacement. The translation simulates the effect of the image moving in a straight line across the frame.
- Scaling. The image is scaled up or down linearly, simulating zoom-in and zoom-out effects. It involves increasing or decreasing the size of the image uniformly from the center, creating the appearance of the camera moving closer to or farther away from the subject.
- Rotation. The image is rotated around its center by a random angle within a specified range, to simulate the effect of the image spinning in place, mimicking a rotational movement of the camera or the object within the scene.

Samples of the synthetic event data are provided in Fig. 4, and we follow [47] for visualizations. When using the training set of the N-ImageNet dataset [20] for training, the performance is 54.36%. By training on the out-of-domain synthetic dataset, we can achieve competitive zero-shot object recognition performance (47.84% vs. 54.36%) on the N-ImageNet dataset [20].

## 4. Experiments

### 4.1. Set-ups

**Training datasets.** To train our methods, we create a synthetic event dataset by curating RGB data from multiple sources, following the pipeline described in Sec. 3.4. In total, there are around 180M RGB images.

**Downstream datasets.** Our method is evaluated on the object and action recognition tasks. For object recognition, N-ImageNet [20], N-Caltech101 [32], CIFAR-10-DVS [23], and N-MINIST [33] datasets are explored. To benchmark our performance on the action recognition tasks, we explore HARDVS [43], DailyAction [27], PAF [30], Bullying10K [11], and UCF101-DVS [4] datasets.

**Baselines.** We compare with two groups of state-of-the-art methods: i) zero-shot frameworks for event data on object and action recognition datasets; ii) dataset-specific methods optimized for different datasets.

**Implementation details.** We train our method for 50 epochs with a batch size of 10240, a layer decay of  $8.5 \times 10^{-1}$ , a gradient clip of 5, a learning rate of  $4 \times 10^{-4}$ , weight decay of  $5 \times 10^{-2}$ , and an AdamW optimizer [29]. The learning rate is cosinely scheduled with a warm epoch of 1. We follow [47] for augmentations. Our event encoder  $f^{\text{evt}}(\cdot)$  uses a ViT-B/16 backbone, initialized from a pre-trained CLIP [40], and we use the same pre-trained CLIP for the image encoder  $f^{\text{img}}(\cdot)$ . All the experiments are conducted under the PyTorch framework [34].

### 4.2. Object Recognition

**Settings.** For methods belonging to the zero-shot learning framework category (*i.e.*, indicated by a tick in the ‘ZS’ column of Tab. 2), we explore their upper bound performance by optimizing text embeddings of corresponding class names. The performance of these variations are presented in rows where the method name superscripted with a ‘up’ symbol. All the zero-shot learning frameworks are based on EVA-02-CLIP-B/16 [40] for fair comparisons.

**Results.** Comparisons with past state-of-the-art methods for object recognition on the N-ImageNet [20], N-Caltech101 [32], CIFAR-10-DVS [23], N-MINIST [33] datasets are given in Tab. 2 (a-d). Our observations are described as follows: i) we have the best zero-shot performance among the groups of zero-shot learning framework on the four datasets; ii) our performance is competitive with the dataset-specific methods on the N-ImageNet [20] and N-Caltech101 [32] datasets. However, it falls behind on the CIFAR-10-DVS [23] and N-MINIST [33] datasets, similar to other zero-shot learning frameworks. This can be attributed to the low-quality of these datasets, *e.g.*, they have a small resolution. Nonetheless, our methods demonstrate greater robustness than other zero-shot frameworks, achieving higher performance; iii) ECLIP and EventCLIP exhibit higher performance than our methods when using optimized text embeddings on the N-MINIST dataset. Considering these methods feed event data represented as frames (differing in event frame representations) to CLIP pre-trained on RGB images, it suggests the RGB image pre-trained CLIP remains competitive in extracting discriminative embeddings for events with simple textures.

### 4.3. Action Recognition.

**Settings.** Applying zero-shot learning frameworks to action recognition extends beyond object semantic understanding, requiring the modeling of subtle semantic variations over time. Besides, the class labels in existing event-based action recognition datasets are often misleading and

Table 2. Comparisons with state-of-the-art data-specific methods and zero-shot frameworks on (a-d) object [20, 23, 32, 33] and (e-i) action [4, 11, 27, 30, 43] recognition using event camera data. The method categories are distinguished by the ‘ZS’ (i.e., zero-shot) column. All methods are evaluated with the accuracy (%) metric. For the zero-shot learning framework, we explore its upper bound performance by using optimized text embeddings for the recognition task, denoted by a superscript <sup>up</sup>.

Method	ZS	Accuracy	Method	ZS	Accuracy	Method	ZS	Accuracy
ECDP [47]	✗	68.31	ECDP [47]	✗	87.66	ECDP [47]	✗	78.00
EventBind [51]	✗	51.40	EventBind [51]	✗	94.08	DEP [19]	✗	78.60
MEM [21]	✗	57.89	MEM [21]	✗	90.10	PSN [12]	✗	85.90
EventBind [51]	✓	3.05	EventBind[51]	✓	67.58	EventCLIP [51]	✓	13.36
EventBind <sup>up</sup> [51]	✓	14.45	EventBind <sup>up</sup> [51]	✓	84.11	EventCLIP <sup>l</sup> [51]	✓	52.46
ECLIP [50]	✓	8.72	ECLIP [50]	✓	53.88	ECLIP [50]	✓	14.61
ECLIP <sup>up</sup> [50]	✓	25.05	ECLIP <sup>up</sup> [50]	✓	74.89	ECLIP <sup>up</sup> [50]	✓	54.43
EventCLIP [44]	✓	4.30	EventCLIP [44]	✓	49.95	EventCLIP [44]	✓	14.05
EventCLIP <sup>up</sup> [44]	✓	17.72	EventCLIP <sup>up</sup> [44]	✓	72.16	EventCLIP <sup>up</sup> [44]	✓	56.97
Ours	✓	47.84	Ours	✓	90.44	Ours	✓	18.48
Ours <sup>up</sup>	✓	57.97	Ours <sup>up</sup>	✓	93.56	Ours <sup>up</sup>	✓	60.00

Method	ZS	Accuracy	Method	ZS	Accuracy	Method	ZS	Accuracy
EventBind [51]	✗	99.27	ESTF [43]	✗	51.22	EARSNN [27]	✗	90.30
E2VID [37]	✗	98.30	TSM [24]	✗	52.63	IRSNN [48]	✗	94.60
RC-GCN [3]	✗	99.00	VideoSwin [28]	✗	51.91	EJESSR [2]	✗	91.03
EventBind [51]	✓	12.89	EventBind [51]	✓	60.66	EventBind [51]	✓	90.33
EventBind <sup>up</sup> [51]	✓	59.07	EventBind <sup>up</sup> [51]	✓	62.53	EventBind <sup>up</sup> [51]	✓	91.17
ECLIP [50]	✓	14.56	ECLIP [50]	✓	37.27	ECLIP [50]	✓	85.64
ECLIP <sup>up</sup> [50]	✓	63.37	ECLIP <sup>up</sup> [50]	✓	39.51	ECLIP <sup>up</sup> [50]	✓	86.19
EventCLIP [44]	✓	11.87	EventCLIP [44]	✓	40.98	EventCLIP [44]	✓	85.08
EventCLIP <sup>up</sup> [44]	✓	65.60	EventCLIP <sup>up</sup> [44]	✓	45.44	EventCLIP <sup>up</sup> [44]	✓	85.64
Ours	✓	25.72	Ours	✓	64.24	Ours	✓	98.62
Ours <sup>up</sup>	✓	59.60	Ours <sup>up</sup>	✓	66.89	Ours <sup>up</sup>	✓	99.17

Method	ZS	Accuracy	Method	ZS	Accuracy	Method	ZS	Accuracy
EARSNN [27]	✗	78.10	R3D [41]	✗	66.80	TIM [38]	✗	63.80
STCA [16]	✗	71.20	SlowFast [15]	✗	69.00	FRMSNN [49]	✗	63.50
AEDCM [45]	✗	55.00	X3D [14]	✗	70.80	3D ResNet [5]	✗	57.90
EventBind [44]	✓	84.72	Eventbind [51]	✓	81.80	EventBind [51]	✓	79.99
EventBind <sup>up</sup> [44]	✓	86.11	Eventbind <sup>up</sup> [51]	✓	81.92	EventBind <sup>up</sup> [51]	✓	81.09
ECLIP [50]	✓	81.94	ECLIP [50]	✓	76.26	ECLIP [50]	✓	74.43
ECLIP <sup>up</sup> [50]	✓	87.50	ECLIP <sup>up</sup> [50]	✓	76.45	ECLIP <sup>up</sup> [50]	✓	74.70
EventCLIP [44]	✓	68.06	EventCLIP [44]	✓	73.24	EventCLIP [44]	✓	71.62
EventCLIP <sup>up</sup> [44]	✓	72.22	EventCLIP <sup>up</sup> [44]	✓	73.52	EventCLIP <sup>up</sup> [44]	✓	72.28
Ours	✓	84.72	Ours	✓	87.87	Ours	✓	81.32
Ours <sup>up</sup>	✓	88.89	Ours <sup>up</sup>	✓	88.18	Ours <sup>up</sup>	✓	81.69

Method	ZS	Accuracy	Method	ZS	Accuracy	Method	ZS	Accuracy
EventBind [51]	✗	99.27	ESTF [43]	✗	51.22	EARSNN [27]	✗	90.30
E2VID [37]	✗	98.30	TSM [24]	✗	52.63	IRSNN [48]	✗	94.60
RC-GCN [3]	✗	99.00	VideoSwin [28]	✗	51.91	EJESSR [2]	✗	91.03
EventBind [51]	✓	12.89	EventBind [51]	✓	60.66	EventBind [51]	✓	90.33
EventBind <sup>up</sup> [51]	✓	59.07	EventBind <sup>up</sup> [51]	✓	62.53	EventBind <sup>up</sup> [51]	✓	91.17
ECLIP [50]	✓	14.56	ECLIP [50]	✓	37.27	ECLIP [50]	✓	85.64
ECLIP <sup>up</sup> [50]	✓	63.37	ECLIP <sup>up</sup> [50]	✓	39.51	ECLIP <sup>up</sup> [50]	✓	86.19
EventCLIP [44]	✓	11.87	EventCLIP [44]	✓	40.98	EventCLIP [44]	✓	85.08
EventCLIP <sup>up</sup> [44]	✓	65.60	EventCLIP <sup>up</sup> [44]	✓	45.44	EventCLIP <sup>up</sup> [44]	✓	85.64
Ours	✓	25.72	Ours	✓	64.24	Ours	✓	98.62
Ours <sup>up</sup>	✓	59.60	Ours <sup>up</sup>	✓	66.89	Ours <sup>up</sup>	✓	99.17

Method	ZS	Accuracy	Method	ZS	Accuracy	Method	ZS	Accuracy
EARSNN [27]	✗	78.10	R3D [41]	✗	66.80	TIM [38]	✗	63.80
STCA [16]	✗	71.20	SlowFast [15]	✗	69.00	FRMSNN [49]	✗	63.50
AEDCM [45]	✗	55.00	X3D [14]	✗	70.80	3D ResNet [5]	✗	57.90
EventBind [44]	✓	84.72	Eventbind [51]	✓	81.80	EventBind [51]	✓	79.99
EventBind <sup>up</sup> [44]	✓	86.11	Eventbind <sup>up</sup> [51]	✓	81.92	EventBind <sup>up</sup> [51]	✓	81.09
ECLIP [50]	✓	81.94	ECLIP [50]	✓	76.26	ECLIP [50]	✓	74.43
ECLIP <sup>up</sup> [50]	✓	87.50	ECLIP <sup>up</sup> [50]	✓	76.45	ECLIP <sup>up</sup> [50]	✓	74.70
EventCLIP [44]	✓	68.06	EventCLIP [44]	✓	73.24	EventCLIP [44]	✓	71.62
EventCLIP <sup>up</sup> [44]	✓	72.22	EventCLIP <sup>up</sup> [44]	✓	73.52	EventCLIP <sup>up</sup> [44]	✓	72.28
Ours	✓	84.72	Ours	✓	87.87	Ours	✓	81.32
Ours <sup>up</sup>	✓	88.89	Ours <sup>up</sup>	✓	88.18	Ours <sup>up</sup>	✓	81.69

Method	ZS	Accuracy	Method	ZS	Accuracy	Method	ZS	Accuracy
EventBind [51]	✗	99.27	ESTF [43]	✗	51.22	EARSNN [27]	✗	90.30
E2VID [37]	✗	98.30	TSM [24]	✗	52.63	IRSNN [48]	✗	94.60
RC-GCN [3]	✗	99.00	VideoSwin [28]	✗	51.91	EJESSR [2]	✗	91.03
EventBind [51]	✓	12.89	EventBind [51]	✓	60.66	EventBind [51]	✓	90.33
EventBind <sup>up</sup> [51]	✓	59.07	EventBind <sup>up</sup> [51]	✓	62.53	EventBind <sup>up</sup> [51]	✓	91.17
ECLIP [50]	✓	14.56	ECLIP [50]	✓	37.27	ECLIP [50]	✓	85.64
ECLIP <sup>up</sup> [50]	✓	63.37	ECLIP <sup>up</sup> [50]	✓	39.51	ECLIP <sup>up</sup> [50]	✓	86.19
EventCLIP [44]	✓	11.87	EventCLIP [44]	✓	40.98	EventCLIP [44]	✓	85.08
EventCLIP <sup>up</sup> [44]	✓	65.60	EventCLIP <sup>up</sup> [44]	✓	45.44	EventCLIP <sup>up</sup> [44]	✓	85.64
Ours	✓	25.72	Ours	✓	64.24	Ours	✓	98.62
Ours <sup>up</sup>	✓	59.60	Ours <sup>up</sup>	✓	66.89	Ours <sup>up</sup>	✓	99.17

uninformative, such as ‘mobile card 1’ and ‘mobile card 2’ from the HARDVS dataset [43]. To address these issues

and explore the performance of all zero-shot frameworks, we train a two-layer ViT to temporally aggregate seman-

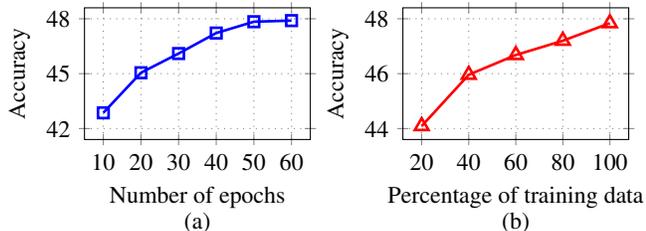


Figure 5. Ablation study on the number of training epochs and percentage of training data. (a) Training epochs are varied from 10 to 60 with a step size of 10. (b) The percentage (%) of synthetic data used in training our method is varied from 20% to 100% with a step size of 20%.

tics from the visual class embeddings of event frames. This model contains approximately 0.35M parameters and can be trained in 10 minutes. Similarly, the upper bound performance is explored by optimizing text embeddings that are indicated in rows of the method name superscripted with the ‘up’ symbol.

**Results.** Tab. 2 (e-i) compares all methods for action recognition on the HARDVS [43], DailyAction [27], PAF [30], Bullying10K [11], and UCF101-DVS [4] datasets. Our findings are threefold: i) all zero-shot learning frameworks serve as strong semantic extractors. Coupled with the two-layer ViT for reasoning temporal dynamics, they generally achieve better scores than the dataset-specific methods in action recognition. Among them, our method showing the best performance; ii) consistent with object recognition task, the optimized text embeddings always lead to a better action recognition performance; iii) our method is significantly better than other zero-shot learning frameworks and dataset-specific methods, on the largest and most complex action recognition dataset, HARDVS [43], demonstrating the effectiveness of our approach.

#### 4.4. Ablations and Discussions

Performance of all the ablation studies are evaluated on the largest event-based object recognition dataset (*i.e.*, the N-ImagetNet dataset).

**Training epochs.** We study our network convergence by showing the zero-shot object recognition performance variations with respect to the number of training epochs in Fig. 5 (a). The training epochs are ranged from 10 to 60, and a step size of 10 is used. A performance saturation is observed after 50 epochs, and motivate us to set the number of training epochs to 50.

**Data scalability.** We validate the effectiveness of our synthetic event data for training. Fig. 5 (b) investigates it by varying the percentage of training data from 20% to 100% with a step size of 20%. Consistently, as more synthetic event data is used, the zero-shot object recognition performance is further improved.

Table 3. Model scalability. We study training with various CLIP image encoder architectures pre-trained from various sources [35, 40]. The performance is evaluated with accuracy (%). ‘-’ denotes unavailability.

Model	Source	Accuracy
ViT-B/32	OpenAI [35]	32.87
	EVA [40]	-
ViT-B/16	OpenAI [35]	40.47
	EVA [40]	47.84
ViT-L/14	OpenAI [35]	51.44
	EVA [40]	57.53
ViT-L/14-336	OpenAI [35]	51.56
	EVA [40]	59.17
ViT-bigE/14+	OpenAI [35]	-
	EVA [40]	62.28

Table 4. Example of integrating our event encoder in place of the one used by LLaVA [26]. The corresponding RGB image of the event frame is shown on the right for reference.

#### Example conversation



User: Caption the image in 80 words.  
Assistant: A bustling city street scene with a crowd of people, a mix of individuals and groups, some walking, others standing, all engaged in their own activities. The atmosphere is vibrant and dynamic, with a sense of movement and energy. The architecture and signage suggest a dense urban environment, possibly in a downtown or central district. The image captures the diversity and liveliness of city life.

**Model scalability.** Tab. 3 studies model scalability in terms of computations and parameters. CLIP based on different image encoder architectures and pre-trained by OpenAI [35] and EVA [13] are investigated. Scaling from 86M parameters (ViT-B) to 4.4B parameters (ViT-bigE) for the CLIP image encoder results in a significant increase in zero-shot performance for our methods. Meanwhile, examining the impact of more computations during inference (*e.g.*, ViT-L/14 vs. ViT-L/14-336), we find a positive correlation with zero-shot performance. Jointly with the data scalability explored in the last section, our method demonstrates the appealing scalability in model parameters and training data.

**Event and text data.** The reader may wonder about the performance of our methods when using synthetic event data and text data. We fine-tune our network using the synthetic event data on the CC3M dataset [6]. However, we observe a decrease in zero-shot performance during train-

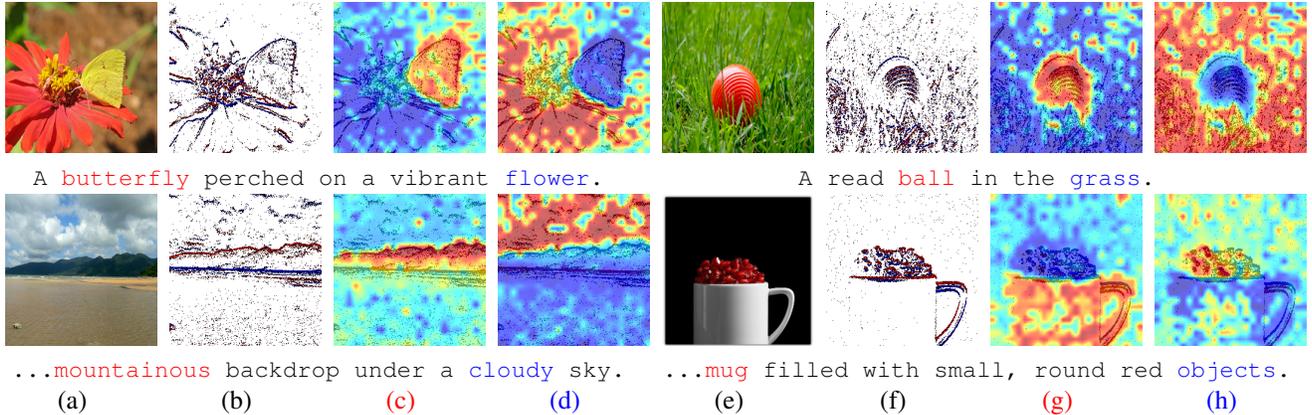


Figure 6. Samples of heatmaps on event frames. The warmer the color, the more important the area. (a)/(e) are RGB images provided as references. (b)/(f) are event frames. Similarly, red and blue indicate positive and negative events. (c)/(d)/(g)/(h) are heatmaps with respect to texts highlighted in the same color, where the text descriptions of the image are provided in the bottom of each row.

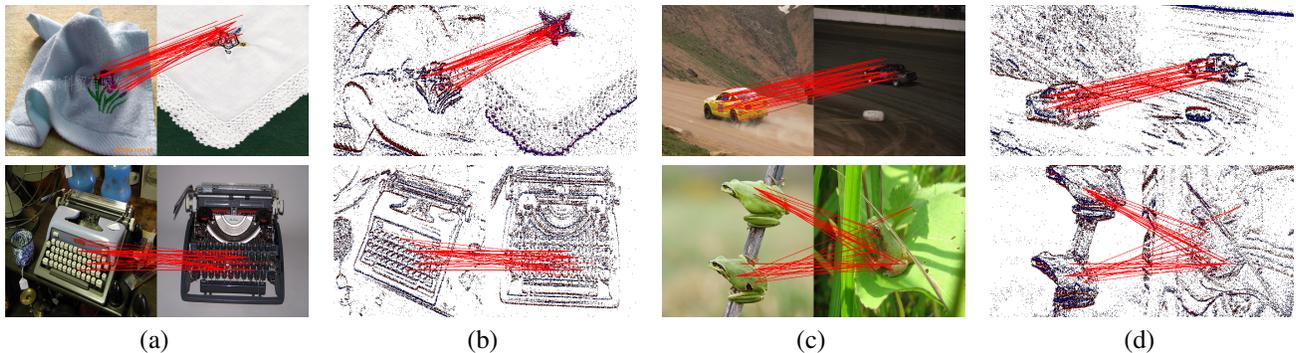


Figure 7. Samples of cross correspondences between two semantically similar event frames. The corresponding RGB images of (b)/(d) are shown in (a)/(c) as references. Note that there are some spatial misalignments between the RGB and event frames. The top cross correspondences extracted from the event frames in (b) and (d) are visualized as red lines.

ing, with an accuracy of only 14.84%, which is significantly lower than the 47.84% accuracy of our method. The performance decrease suggests the effectiveness of our approach.

**Model analysis.** We evaluate our model from three perspectives: i) Tab. 4 demonstrates conversations with a vision-language model, where our event encoder is integrated as a replacement for the CLIP image encoder used in LLaVA [26]; ii) Fig. 6 visualizes heatmaps of event data in relation to corresponding texts; iii) Fig. 7 extracts the sparse cross-correspondences between two semantically similar event frames. Overall, our model exhibits a robust ability to understand semantic content. For instance, in Fig. 6, the model trained on event and RGB data pairs effectively captures meaningful textual semantics. Specifically, on the right side of the second row, the model accurately identifies the last word, ‘objects’, described as ‘small’ and ‘round’, as referring to pomegranates. This strong semantic comprehension ability underpins our performance in tasks such as object and action recognition.

**Limitations.** While progressing in zero-shot object recognition, a gap remains in developing a generalist foun-

dation model for event camera data. However, as shown by [9, 26], a versatile pre-trained CLIP image encoder is a key component for such models, which this paper investigates.

## 5. Conclusion and Broader Impact

We propose a framework to unlock zero-shot object recognition performance within the event camera data community. Our method is trained with event and RGB data pairs. To address the scarcity of these data pairs, we study a pipeline to synthesize event data from still RGB images. By analyzing the limitations of existing approaches, our key insights are to perform the scalar-wise semantic modulation during training. Experiments on standard object and action recognition benchmark datasets extensively demonstrate the superiority of our approach compared to previous state-of-the-art methods.

**Broader impact.** We believe our method is promising to be applied in training a generative event-vision-language foundation model. We hope it inspires future work.

**Acknowledgements.** Liyuan Pan’s 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] Arnon Amir, Brian Taba, David J. Berg, Timothy Melano, Jeffrey L. McKinstry, Carmelo di Nolfo, Tapan K. Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael DeBole, Steven K. Esser, Tobi Delbrück, Myron Flickner, and Dharmendra S. Modha. A low power, fully event-based gesture recognition system. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7388–7397. IEEE Computer Society, 2017. 3
- [2] Sami Barchid, José Mennesson, and Chaabane Djeraba. Exploring joint embedding architectures and data augmentations for self-supervised representation learning in event-based vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 3903–3912. IEEE, 2023. 6
- [3] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based object classification for neuromorphic vision sensing. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 491–501. IEEE, 2019. 3, 6
- [4] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based spatial-temporal feature learning for neuromorphic vision sensing. *CoRR*, abs/1910.03579, 2019. 1, 3, 5, 6, 7
- [5] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based spatio-temporal feature learning for neuromorphic vision sensing. *IEEE Trans. Image Process.*, 29:9084–9098, 2020. 6
- [6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3558–3568. Computer Vision Foundation / IEEE, 2021. 7
- [7] Xinlei Chen\*, Saining Xie\*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 3
- [8] Hoonhee Cho, Hyeonseong Kim, Yujeong Chae, and Kuk-Jin Yoon. Label-free event-based object recognition via joint learning with image reconstruction from events. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19809–19820. IEEE, 2023. 2, 3
- [9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 8
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 2
- [11] Yiting Dong, Yang Li, Dongcheng Zhao, Guobin Shen, and Yi Zeng. Bullying10k: A large-scale neuromorphic dataset towards privacy-preserving bullying recognition. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 1, 5, 6, 7
- [12] Wei Fang, Zhaofei Yu, Zhaokun Zhou, Ding Chen, Yanqi Chen, Zhengyu Ma, Timothée Masquelier, and Yonghong Tian. Parallel spiking neurons with high efficiency and ability to learn long-term dependencies. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 6
- [13] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: exploring the limits of masked visual representation learning at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19358–19369. IEEE, 2023. 2, 7
- [14] Christoph Feichtenhofer. X3D: expanding architectures for efficient video recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 200–210. Computer Vision Foundation / IEEE, 2020. 6
- [15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6201–6210. IEEE, 2019. 6
- [16] Pengjie Gu, Rong Xiao, Gang Pan, and Huajin Tang. STCA: spatio-temporal credit assignment with delayed feedback in deep spiking neural networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 1366–1372. ijcai.org, 2019. 6
- [17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022. 3
- [18] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbrück. v2e: From video frames to realistic DVS events. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 1312–1321. Computer Vision Foundation / IEEE, 2021. 5
- [19] Zhenpeng Huang, Chao Li, Hao Chen, Yongjian Deng, Yifeng Geng, and Limin Wang. Data-efficient event camera pre-training via disentangled masked modeling, 2024. 6

- [20] 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 (ICCV)*, pages 2146–2156, 2021. 1, 2, 3, 4, 5, 6
- [21] Simon Klenk, David Bonello, Lukas Koestler, Nikita Araslanov, and Daniel Cremers. Masked event modeling: Self-supervised pretraining for event cameras. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pages 2367–2377. IEEE, 2024. 3, 6
- [22] Lingdong Kong, Youquan Liu, Lai Xing Ng, Benoit R. Cottereau, and Wei Tsang Ooi. Openess: Event-based semantic scene understanding with open vocabularies. *CoRR*, abs/2405.05259, 2024. 1, 2
- [23] Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: An event-stream dataset for object classification. *Frontiers in Neuroscience*, 11, 2017. 1, 3, 5, 6
- [24] Ji Lin, Chuang Gan, and Song Han. TSM: temporal shift module for efficient video understanding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7082–7092. IEEE, 2019. 6
- [25] Haotian Liu, Guang Chen, Sanqing Qu, Yanping Zhang, Zhi-jun Li, Alois Knoll, and Changjun Jiang. Tma: Temporal motion aggregation for event-based optical flow. In *ICCV*, 2023. 2, 4
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 7, 8
- [27] Qianhui Liu, Dong Xing, Huajin Tang, De Ma, and Gang Pan. Event-based action recognition using motion information and spiking neural networks. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1743–1749. ijcai.org, 2021. 1, 3, 5, 6, 7
- [28] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *CoRR*, abs/2106.13230, 2021. 6
- [29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 5
- [30] Shu Miao, Guang Chen, Xiangyu Ning, Yang Zi, Kejia Ren, Zhenshan Bing, and Alois C. Knoll. Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. *Frontiers Neurobotics*, 13:38, 2019. 1, 3, 5, 6, 7
- [31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafranec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *CoRR*, abs/2304.07193, 2023. 3
- [32] Garrick Orchard, Ajinkya Jayawant, Gregory Cohen, and Nitish V. Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *CoRR*, abs/1507.07629, 2015. 1, 3, 5, 6
- [33] Garrick Orchard, Ajinkya Jayawant, Gregory Cohen, and Nitish V. Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *CoRR*, abs/1507.07629, 2015. 1, 3, 5, 6
- [34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. 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 *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019. 5
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 1, 2, 7
- [36] Bharath Ramesh, Andrés Ussa, Luca Vedova, Hong Yang, and Garrick Orchard. Low-power dynamic object detection and classification with freely moving event cameras. *Frontiers in Neuroscience*, 14, 2020. 3
- [37] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3857–3866. Computer Vision Foundation / IEEE, 2019. 6
- [38] Sicheng Shen, Dongcheng Zhao, Guobin Shen, and Yi Zeng. TIM: an efficient temporal interaction module for spiking transformer. *CoRR*, abs/2401.11687, 2024. 6
- [39] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. HATS: histograms of averaged time surfaces for robust event-based object classification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1731–1740. Computer Vision Foundation / IEEE Computer Society, 2018. 3
- [40] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389, 2023. 1, 5, 7
- [41] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri. Convnet architecture search for spatiotemporal feature learning. *CoRR*, abs/1708.05038, 2017. 6
- [42] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 3

- [43] Xiao Wang, Zongzhen Wu, Bo Jiang, Zhimin Bao, Lin Zhu, Guoqi Li, Yaowei Wang, and Yonghong Tian. Hardvs: Revisiting human activity recognition with dynamic vision sensors, 2022. [1](#), [3](#), [5](#), [6](#), [7](#)
- [44] Ziyi Wu, Xudong Liu, and Igor Gilitschenski. Eventclip: Adapting CLIP for event-based object recognition. *CoRR*, abs/2306.06354, 2023. [1](#), [2](#), [3](#), [6](#)
- [45] Rong Xiao, Huajin Tang, Yuhao Ma, Rui Yan, and Garrick Orchard. An event-driven categorization model for AER image sensors using multispikes encoding and learning. *IEEE Trans. Neural Networks Learn. Syst.*, 31(9):3649–3657, 2020. [6](#)
- [46] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data dense pre-training. *CoRR*, abs/2311.11533, 2023. [1](#), [3](#)
- [47] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10699–10709, 2023. [1](#), [3](#), [5](#), [6](#)
- [48] Man Yao, Jiakui Hu, Guangshe Zhao, Yaoyuan Wang, Ziyang Zhang, Bo Xu, and Guoqi Li. Inherent redundancy in spiking neural networks. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 16878–16888. IEEE, 2023. [6](#)
- [49] Man Yao, Hengyu Zhang, Guangshe Zhao, Xiyu Zhang, Dingheng Wang, Gang Cao, and Guoqi Li. Sparser spiking activity can be better: Feature refine-and-mask spiking neural network for event-based visual recognition. *Neural Networks*, 166:410–423, 2023. [6](#)
- [50] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. E-CLIP: towards label-efficient event-based open-world understanding by CLIP. *CoRR*, abs/2308.03135, 2023. [1](#), [2](#), [3](#), [6](#)
- [51] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Eventbind: Learning a unified representation to bind them all for event-based open-world understanding, 2024. [1](#), [2](#), [3](#), [6](#)