

CEIA: CLIP-Based Event-Image Alignment for Open-World Event-Based Understanding

Wenhao Xu, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong

University of Science and Technology of China

Abstract. We present CEIA, an effective framework for open-world event-based understanding. Currently training a large event-text model still poses a huge challenge due to the shortage of paired event-text data. In response to this challenge, CEIA learns to align event and image data as an alternative instead of directly aligning event and text data. Specifically, we leverage the rich event-image datasets to learn an event embedding space aligned with the image space of CLIP through contrastive learning. In this way, event and text data are naturally aligned via using image data as a bridge. Particularly, CEIA offers two distinct advantages. First, it allows us to take full advantage of the existing event-image datasets to make up the shortage of large-scale event-text datasets. Second, leveraging more training data, it also exhibits the flexibility to boost performance, ensuring scalable capability. In highlighting the versatility of our framework, we make extensive evaluations through a diverse range of event-based multi-modal applications, such as object recognition, event-image retrieval, event-text retrieval, and domain adaptation. The outcomes demonstrate CEIA’s distinct zero-shot superiority over existing methods on these applications.

Keywords: Event-Based Understanding · Zero-Shot · Multi-Modal

1 Introduction

Event cameras are sensors that asynchronously measure the intensity changes at each pixel independently with microsecond temporal resolution [11]. Compared to conventional frame cameras, event cameras exhibit several exceptional advantages. They have a very high dynamic range, are immune to motion blur, and provide measurements with a microsecond-level temporal resolution. These inherent advantages have sparked considerable interest in event cameras, notably for computer vision applications such as autonomous navigation [21], robotics [13], and virtual reality (VR) [28].

Despite the superiority of event cameras, event-based algorithms are still in their infancy, facing two major issues: the shortage of large-scale datasets and the failure of modeling new data distributions in the real world. Consequently, it is imperative to explore the zero-shot event-based algorithms. Very recently, several works [45, 55] have explored how to transfer impressive zero-shot knowledge

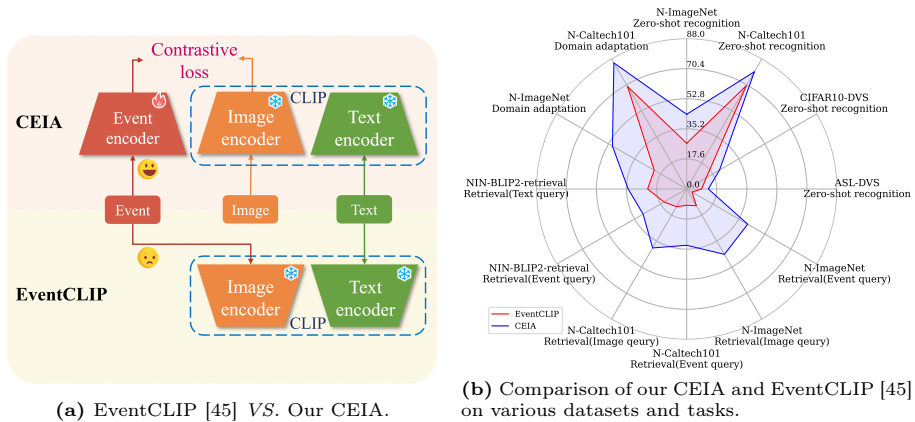


Fig. 1: (a) Compared with EventCLIP [45] that directly utilizes the frozen CLIP’s image encoder, our CEIA learns an event encoder to alleviate the event-image modality disparity. (b) Comparison of our CEIA and EventCLIP [45] on various datasets and tasks. For zero-shot recognition and domain adaptation, we report Acc1 (%), while for event-image retrieval and event-text retrieval, we report R@1 (%) [25].

from CLIP [37] to event-based vision. EventCLIP [45] demonstrated the feasibility of improving event-based zero-shot capability by first transforming events into frames and then directly utilizing frozen CLIP to extract event features. However, the image encoder of CLIP is primarily trained on natural images, resulting in a significant domain gap between images and the transformed frames. Therefore, the performance is severely impeded. To address this shortcoming, in this paper, we propose CEIA, an effective framework to adapt CLIP to event data while accommodating a wide range of open-world event-based understanding tasks.

CEIA achieves its goal by learning an individual event encoder through cross-modal contrastive learning, instead of directly utilizing the frozen image encoder like EventCLIP. The differences between EventCLIP and CEIA are depicted in Fig. 1a. In particular, we observe that, unlike 3D point clouds [14] or depth maps, event data have a notable characteristic: they are often accompanied by available paired image data. This accessibility is largely thanks to the widespread use of dynamic and active-pixel vision sensors (DAVIS) [2, 4], which can simultaneously capture pixel-wise images and event data. Leveraging this advantage, we provide a novel perspective of training an event encoder using abundant paired event-image data instead of directly conducting event-text alignment, thus bypassing the shortage of large-scale paired event-text data. Instead of full finetuning the event encoder, we introduce a simple yet highly-efficient training strategy based on the LoRA [18] technique to focus on relating the event and image modalities, meanwhile preserving the highly-robust zero-shot ability provided by CLIP. In this way, CEIA can learn an event embedding space aligned with the image embedding space of frozen CLIP. Notably, the image space is already aligned

with the text space during pretraining by CLIP. Consequently, event and text data are also naturally aligned by using image data as a bridge. By this way, CEIA can not only enhance open-world event-text understanding but also open the door to more event-based multi-modal understanding tasks [6, 29, 44, 46].

In highlighting the versatility of CEIA, we make extensive evaluations through a diverse range of multi-modal understanding tasks. CEIA, designed to strike a unified embedding space for aligning event, image, and text data, can be smoothly applied to object recognition, event-image retrieval, event-text retrieval, and domain adaptation [32, 42]. The experimental outcomes demonstrate that the state-of-the-art zero-shot performance can be achieved by CEIA over the existing methods, which further spotlights CEIA’s transferability and versatility. Additionally, we observe that, leveraging more training data, CEIA also exhibits the flexibility to yield a significant performance boost, ensuring the scalable capability. Through these extensive experimental evaluations on four applications, as shown in Fig. 1b, we not only confirm the exceptional functionality of event, image, and text alignment of CEIA, but also underscore the comprehensive application capabilities of CEIA. We believe that CEIA stands as a robust and effective framework for open-world event-based multi-modal understanding.

In summary, CEIA presents three main contributions: (i) an effective framework to provide a novel perspective of learning to align event and image data as an alternative, thus bypassing the shortage of event-text datasets. (ii) a simple yet highly-efficient strategy for training the event encoder with the LoRA technique, meanwhile preserving the CLIP’s powerful robustness. (iii) state-of-the-art results on four event-based multi-modal downstream tasks, including zero-shot and few-shot object recognition, event-text retrieval, event-image retrieval, and domain adaptation.

2 Related Work

2.1 Transferring CLIP

In the image-based vision, pretrained Visual Language Models like CLIP [37], ALIGN [19], and Florence [50] demonstrate very impressive zero-shot transfer and generalization capabilities. Subsequently, a large number of follow-up works have been proposed to transfer the pretrained CLIP to more downstream tasks. For example, PointCLIP [51] transforms 3D point clouds into a set of depth maps for zero-shot 3D object recognition, while DenseCLIP [38] converts the original image-text matching to pixel-text matching to guide the learning of dense prediction models. X-CLIP [33] proposes a novel cross-frame attention mechanism to effectively expand CLIP to the video domain.

Recently, some works have applied Visual Language Models to event-based vision, demonstrating promising results. Two works closely related to ours are EventCLIP [45] and E-CLIP [55]. Similar to PointCLIP, EventCLIP first transforms events into 2D frames and then uses frozen CLIP directly for zero-shot event object recognition. Following EventCLIP, E-CLIP focuses on advancing

few-shot and standard object recognition. It introduces a novel event encoder for event temporal modeling and presents a triple contrastive alignment module to enable efficient knowledge transfer. In contrast, instead of directly utilizing frozen CLIP, we leverage existing abundant event-image datasets to adapt CLIP to event-based zero-shot tasks.

2.2 Multi-Modal Learning

With the availability of large-scale multi-modal datasets, an increasing number of multi-modal foundation models have emerged. Some representative models are driving multi-modal learning, which has marked a significant advancement in AI evolution. For example, CLIP [37] demonstrates impressive zero-shot object recognition performance, while BLIP-2 [27] exhibits capabilities approaching human-level performance in visual dialog, visual knowledge reasoning, and personalized image-to-text generation. Furthermore, Stable Diffusion [40] can generate realistic and accurate images based on given text conditions. Instruct-Pix2Pix [5] can execute diverse image edits following human-written instructions, including object replacement, style modification, setting changes, and adjustments to the artistic medium.

These advancements motivate us to explore event-based multi-modal tasks. In this paper, we study two main directions. One is event-text understanding, including zero-shot learning and event-text retrieval, while the other is event-image understanding, involving event-image retrieval and domain adaptation. Our future work will focus on generalizing CEIA for wider multi-modal tasks, such as event-assisted video frame interpolation [36, 43, 46, 49] and event-assisted motion deblurring [29, 44, 53, 54].

3 Method

3.1 CLIP Preliminaries

CLIP [37] is a visual-text pre-training method for image and text matching. Conceptually, CLIP consists of two encoders: an image encoder $\Phi_{image}(\cdot; \theta_0)$ for extracting visual features and a text encoder $\Phi_{text}(\cdot; \theta_1)$ for extracting text features. During training, CLIP utilizes 400 million training image-text pairs collected from the internet and employs a contrastive loss to learn a unified embedding space for accommodating image and text data. Specifically, given a set of image-text pairs $\{\mathbf{x}^{image}, \mathbf{x}^{text}\}$, CLIP is trained to search optimized parameters θ_0 and θ_1 to approach

$$\Phi_{image}(\mathbf{x}^{image}; \theta_0) = \Phi_{text}(\mathbf{x}^{text}; \theta_1). \quad (1)$$

Note that we use “=” to denote the alignment in the whole paper. Leveraging the large-scale image-text dataset, CLIP demonstrates promising zero-shot performance for many downstream tasks, ensuring the incorporation of a huge range of visual concepts.

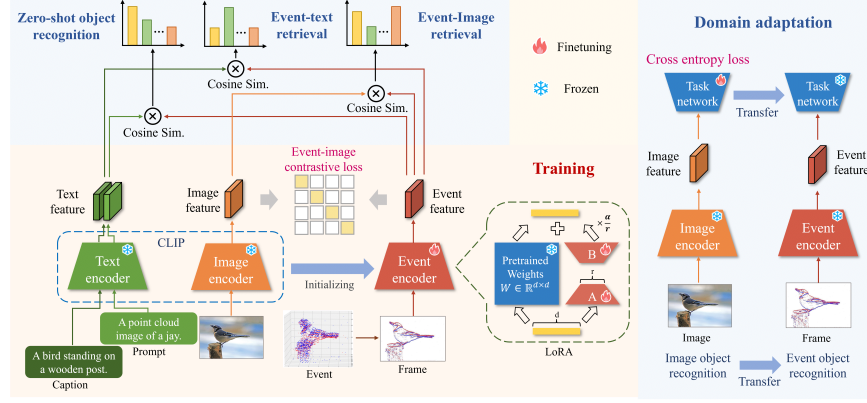


Fig. 2: Overview of CEIA, which consists of a learnable event encoder, a frozen image encoder, and a frozen text encoder. We initialize the event encoder with CLIP’s image encoder and finetune it using the LoRA [18] technique. We align the event embedding space and image embedding space through contrastive learning. In highlighting the versatility of CEIA, we make evaluations on four applications: object recognition, event-text retrieval, event-image retrieval, and domain adaptation.

3.2 The CEIA Framework

In particular, open-world event-based multi-modal understanding still remains under-explored. Our goal is to transfer the zero-shot capability of CLIP into the event-based vision. To this end, two challenges need to be addressed. First, intuitively, one way to achieve open-world event-based understanding is to train a large event-text model. Nevertheless, it is severely impeded due to the shortage of large-scale paired event-text data. Second, compared with natural images, the event data, captured by detecting the intensity changes, is essentially a kind of spatial-temporal data. Therefore, the big modality disparity makes it difficult to directly apply the image encoder of CLIP to event data.

In response to these two challenges, CEIA makes two key modifications. First, CEIA provides a novel perspective of focusing on learning to align event and image data instead of conducting event-text alignment, thus bypassing the shortage of large-scale paired event-text data. Second, CEIA learns an individual event encoder to alleviate the event-image modality disparity instead of directly utilizing the frozen image encoder like EventCLIP [45]. In the following, we will formally introduce the method.

Overview. Fig. 2 shows an overview of CEIA, which is composed of a frozen image encoder $\Phi_{image}(\cdot; \theta_0)$, a frozen text encoder $\Phi_{text}(\cdot; \theta_1)$ and a learnable event encoder $\Phi_{event}(\cdot; \theta_2)$. Given a triple set of image-event-text pairs $\{\mathbf{x}^{event}, \mathbf{x}^{image}, \mathbf{x}^{text}\}$, CEIA learns to search a desirable parameter θ_2 , which meets the following requirement:

$$\Phi_{event}(\mathbf{x}^{event}; \theta_2) = \Phi_{image}(\mathbf{x}^{image}; \theta_0) \quad (2)$$

Notably, CLIP has already provided the powerful image-text alignment as shown in Eq. (1). Consequently, through combining Eq. (1) and Eq. (2), we can align event and text data by regarding $\Phi_{image}(\mathbf{x}^{image}; \theta_0)$ as a bridge

$$\Phi_{event}(\mathbf{x}^{event}; \theta_2) = \Phi_{text}(\mathbf{x}^{text}; \theta_1). \quad (3)$$

Event Encoder. Following existing research [22, 48, 55], we selected the Vision Transformer [8], a reliable and widely-used model, as our event encoder. Leveraging the unified encoder architecture, we propose initializing the event encoder with CLIP’s image encoder and then finetuning it, instead of training it from scratch. This initialization transfers spatial prior knowledge from images to events, accelerating the training process and enhancing the data efficiency of CEIA. In our experiments, this initialization proved not only beneficial but also essential. Since the training data is still too limited for cross-modal alignment, we attempted to train the event encoder from scratch, but failed.

Event Representations. We explored various event representations and determined that the red-blue color map, commonly used for visualizing events, is the most effective. This choice minimizes the difference between the event representation and the natural images used by CLIP, thereby simplifying cross-modal alignment.

LoRA-Based Finetuning. Intuitively, one simple way to learn an event encoder is full finetuning. However, it will destroy the original CLIP’s weights, which brings the inferior zero-shot capability. Recently, LoRA [18] stands out as one of the best parameter-efficient transfer learning methods, which has been widely adopted to finetune many LLMs. Specifically, LoRA [18] shows that the pretrained models can still learn efficiently even when projected into a smaller subspace. For each pretrained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, we can replace its update with a low-rank decomposition $\Delta W = BA$, where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$. Note that W_0 is frozen, while A and B are trainable. For the original forward pass $h = W_0x$, the modified forward pass is:

$$h = W_0x + \Delta Wx = W_0x + \frac{\alpha}{r}BAx \quad (4)$$

where α is a hyperparameter used to adjust the influence of the new parameters. LoRA-based finetuning provides three key advantages for CEIA: 1) It avoids catastrophic forgetting, thus preserving CLIP’s strong generalization and zero-shot capabilities. 2) It prevents overfitting to the limited training data. 3) It significantly reduces training time and memory costs.

Event-Image Contrastive Learning. The objective of training our event encoder $\Phi_{event}(\cdot; \theta_2)$ is to minimize the distance between the frames transformed from events and images in the same pair, while maximizing the distance of others. We draw the inspiration from many methods [15, 17, 37, 47, 55], which advocates the utilization of multi-modal contrastive learning. Specifically, given a set of event-image pairs $\{\mathbf{x}_i^{event}, \mathbf{x}_i^{image}\}_{i=1}^N$, we encode them into normalized embeddings: $f_i^{event} = \Phi_{event}(\mathbf{x}_i^{event}; \theta_2)$ and $f_i^{image} = \Phi_{image}(\mathbf{x}_i^{image}, \theta_0)$. Denoting M_1

and M_2 are two modalities, the InfoNCE [34] loss can be formulated as

$$L(M_1, M_2) = -\log \frac{\exp(f_i^{M_1} \cdot (f_i^{M_2})^T / \tau)}{\exp(f_i^{M_1} \cdot (f_i^{M_2})^T / \tau) + \sum_{j \neq i} \exp(f_i^{M_1} \cdot (f_j^{M_2})^T / \tau)}, \quad (5)$$

where τ is a learnable temperature parameter to control the smoothness of the softmax distribution. Following CLIP [37], we consider every example $j \neq i$ in the mini-batch as a negative. Finally, the weights of the event encoder θ_2 is optimized by minimizing a symmetric InfoNCE loss

$$L_{final} = L(event, image) + L(image, event). \quad (6)$$

Through event-image contrastive learning, we can align representations of event, image, and text modalities into the same embedding space. In the following, we will elaborate on the details about how to extend CEIA to open-world event-based multi-modal applications.

3.3 Event-Based Multi-Modal Applications

Object Recognition. Zero-shot object recognition aims to classify objects that are not included in the training dataset. As shown in Eq. (3), CEIA has achieved event-text alignment in an indirect manner. Through this event-text alignment, CEIA enables zero-shot event-based object recognition. Specifically, we first construct text prompts by inserting the class names of new objects into predefined templates (e.g., “image of a [CLASS]”). Then, we extract their textual features W_t by $\Phi_{text}(\cdot; \theta_2)$. Since each row vector in W_t encodes class knowledge, W_t can naturally function as the zero-shot event classifier. Meanwhile, we utilize $\Phi_{event}(\cdot; \theta_1)$ to extract the event features f_i^{event} from the input events. Finally, the predicted probabilities for K classes are computed via the classifier as follows:

$$logits_i = f_i^{event} W_t^T; p_i = \text{softmax}(logits_i). \quad (7)$$

Similarly, Eq. (7) can be also utilized for few-shot object recognition.

Event-Image/Event-Text Retrieval. Event-image retrieval refers to the task of searching for the most related image in a large-scale image dataset based on a given event, or vice versa. For instance, when given an image query \mathbf{x}_q^{image} , we first extract its image feature f_q^{image} using $\Phi_{image}(\cdot; \theta_0)$. Then, we feed forward all event examples $\{\mathbf{x}_j^{event}\}_{j=1}^N$ into $\Phi_{event}(\cdot; \theta_2)$ to obtain $\{f_j^{event}\}_{j=1}^N$. Subsequently, we calculate their cosine similarity and retrieve the most related event $\mathbf{x}_{j^*}^{event}$ with the highest similarity score:

$$j^* = \underset{j}{\operatorname{argmax}} \left(\frac{f_q^{image} (f_j^{event})^T}{\|f_q^{image}\| \|f_j^{event}\|} \right) \quad (8)$$

For event-text retrieval, we calculate the similarity score between event features and text features and select the item with the highest similarity score.

Domain Adaptation. Domain adaptation [9, 32, 42] aims to transfer tasks from a labeled source domain (images) to a target domain (events). It can leverage existing image datasets to train models, thereby overcoming the lack of high-quality labeled event datasets. Specifically, As depicted in Fig. 2, we conducted domain adaptation experiments on object recognition. Formally, denote f^{image} and l as the image feature extracted by $\Phi_{image}(\cdot; \theta_0)$ and the available label, respectively. We train a task network $T(\cdot; \theta_4)$, whose weights θ_4 are optimized by minimizing the commonly-used soft-max cross-entropy loss:

$$logits = T(f^{image}; \theta_4); L_{image} = CrossEntropy(logits, l). \quad (9)$$

Subsequently, we directly apply the trained task network $T(\cdot; \theta_4)$ to the event domain to generate predictions:

$$pred = T(f^{event}; \theta_4). \quad (10)$$

As shown in Eq. (2), CEIA has already aligned event and image data, which ensures the transferability and applicability of the network $T(\cdot; \theta_4)$ when applied for event data.

4 Experiments

4.1 Dataset Preparation

N-ImageNet. N-ImageNet [20] is built by moving an event camera in front of an LCD monitor which displays images from ImageNet [7]. We leverage the event-image pairs from N-ImageNet [20] and ImageNet-1K [7] for training. Similar to ImageNet-1K, N-ImageNet contains 1.78 million event streams belonging to 1,000 classes. For training, we split N-ImageNet to construct two subset datasets: the Small dataset includes 129,393 event streams belonging to the first 100 classes and the Large dataset includes 638,878 belonging to the first 500 classes. We call the method “X” trained on Small and Large datasets as “X-S” and “X-L”, respectively. We use the Small and Large datasets to explore the scalable capability of our approach. We utilize the official splitting to obtain the training and test datasets.

N-Caltech101. Similar to N-ImageNet [20], N-Caltech101 [35] is built by moving a 180×240 resolution ATIS event camera in front of a monitor displaying still images from Caltech101 [10]. It contains 8,246 samples, each with a duration of 300 ms, belonging to 101 classes. We adopt the same splitting strategy as EST [12] to obtain the training and test datasets.

CIFAR10-DVS. Unlike N-Caltech101 [35] and N-ImageNet [20], CIFAR10-DVS [26] is created through repeating smooth movements of images on an LCD monitor in front of a DVS camera. This process converts the popular CIFAR-10 [23] dataset into 10,000 event streams across 10 different classes. We randomly allocate 4,000 samples for the test set and 6,000 samples for the training set.

ASL-DVS. ASL-DVS [3] is a relatively complex dataset containing the second largest number of labeled examples. It contains 24 classes corresponding to 24

letters (A-Y, excluding J) of the American Sign Language. For each letter, 4,200 samples are collected by capturing real-world events. Each sample spans approximately 100 milliseconds. We randomly select 1,000 samples for the test set and 3,200 samples for the training set.

NIN-Prompt/NIN-BLIP2/NIN-BLIP2-retrieval. Considering the shortage of currently available large-scale event-text datasets, we make the first attempt to build two kinds of event-text datasets based on N-ImageNet for training, denoted as “NIN-Prompt” and “NIN-BLIP2”. Specifically, for “NIN-Prompt”, we first create prompts by placing the class names of events into the template “A point cloud image of [CLASS]”. We then use these prompts as captions for corresponding events. For “NIN-BLIP2”, we utilize BLIP2 [27] with the frozen LLM OPT [52] to conduct zero-shot image captioning, generating high-quality captions for the images from ImageNet. Subsequently, we pair them with events from N-ImageNet to construct the dataset. Furthermore, we create a test dataset, named “NIN-BLIP2-retrieval”, for evaluating event-text retrieval. However, the captions generated from images belonging to the same class are too similar, which may correspond to multiple events. To mitigate this issue, we selectively sample only five images from each class to generate captions, constructing a test set containing 2,500 event-text pairs.

4.2 Implementation Details

We initialize our event encoder with the ViT-L/14 [8] image encoder of CLIP. The AdamW [31] optimizer and a cosine schedule warm-up learning rate schedule [30] are adopted for training. For LoRA-based finetuning [18], we set the peak learning rate to 5×10^{-4} and the weight decay to 1×10^{-2} . For full finetuning, we set the peak learning rate to 1×10^{-7} and the weight decay to 1×10^{-1} . The training batch size is set to 128 for all experiments. Additionally, we conduct prompt engineering and create task-relevant templates for each dataset. Specifically, we adopt “A point cloud image representing the American Sign Language letter [CLASS]” for ASL-DVS, “Image of a [CLASS]” for N-Caltech101, and “A point cloud image of a [CLASS]” for CIFAR10-DVS and N-ImageNet.

4.3 Baselines

We compare CEIA with the current state-of-the-art event-based zero-shot method, EventCLIP [45]. Additionally, we combine the pre-trained event-based video reconstruction network E2VID [39] with the frozen CLIP to construct another simple zero-shot method, which is denoted as “E2VID-CLIP”.

Moreover, leveraging our building event-text datasets NIN-Prompt and NIN-BLIP2, we are able to directly train a CLIP-based event-text alignment model, called “CETA”. We denote “CETA” trained on such two datasets as “CETA-Prompt” and “CETA-BLIP2”, respectively.

Table 1: Quantitative results of zero-shot object recognition.

Method	In-Distribution		Out-of-Distribution					
	N-ImageNet [20]		N-Caltech101 [35]		CIFAR10-DVS [26]		ASL-DVS [3]	
	Acc1	Acc5	Acc1	Acc5	Acc1	Acc5	Acc1	Acc5
EventCLIP [45]	26.72	39.39	69.73	85.93	13.23	56.17	8.72	25.97
E2VID-CLIP	13.68	27.44	82.53	93.62	13.85	55.57	8.43	25.82
CETA-Prompt-S	29.35	50.73	70.93	86.61	14.24	59.35	8.87	28.23
CETA-BLIP2-S	33.86	57.53	75.01	88.68	16.27	65.49	7.53	24.88
CEIA-S	<u>37.25</u>	<u>61.60</u>	72.31	86.50	<u>18.40</u>	<u>65.75</u>	<u>12.11</u>	32.62
CEIA-L	43.68	68.78	<u>79.20</u>	<u>90.80</u>	22.20	69.07	12.67	<u>31.20</u>

Table 2: Quantitative results of few-shot object recognition. Acc1 (%) is reported.

Datasets	N-ImageNet [20]					N-Caltech101 [35]				
Data per Class	1	2	5	10	20	1	2	5	10	20
Sorted Time Surface [1]	1.24	2.19	4.26	7.53	12.81	27.80	31.99	54.85	65.94	75.47
DiST [20]	1.16	1.75	4.22	7.65	13.07	26.42	28.14	53.48	65.71	73.92
EventCLIP [45]	<u>29.41</u>	<u>31.14</u>	<u>32.56</u>	<u>33.08</u>	<u>36.40</u>	<u>75.82</u>	<u>78.86</u>	<u>83.57</u>	<u>87.42</u>	<u>90.41</u>
CEIA-L	44.77	46.07	49.58	51.40	53.32	84.46	87.16	89.28	90.71	92.14

4.4 Object Recognition

Metrics. We evaluate the performance of object recognition in terms of the common top-1 accuracy (Acc1) and top-5 accuracy (Acc5) [16, 41].

Zero-Shot Results. Event-based zero-shot object recognition is a challenging task because the classes in the test set are unseen to the model during training. We report the in-distribution and out-of-distribution results in Tab. 1. The experimental results indicate that our CEIA consistently outperforms the state-of-the-art baselines across all datasets. For instance, on N-ImageNet and N-Caltech101, CEIA-L achieves improvements of 16.96% and 9.47% in top-1 accuracy compared with EventCLIP, respectively. These improvements highlight the effectiveness of our CEIA for open-world event-based understanding. Although E2VID-CLIP achieves better results than ours on N-Caltech101, its complex reconstruction network introduces significant inference latency.

Besides, we notice that CETA-BLIP2 achieves better zero-shot results than CETA-Prompt, which can be attributed to the reason that BLIP2 is able to generate more accurate captions compared with the simple prompt template. However, the event-text alignment method CETA-BLIP2 (CETA-Prompt) exhibits inferior results compared with our event-image alignment method CEIA, highlighting the effectiveness of our event-image alignment strategy compared with the direct event-text alignment strategy.

Few-Shot Results. We consider a general N-shot setting, i.e., N examples are randomly sampled from each class for training. We compare our CEIA with the current state-of-the-art few-shot classifier, EventCLIP [45]. In addition, we also compare with some representative methods without CLIP, namely, Sorted Time Surface [1] and DiST [20]. Notice that, we follow the original papers and use ResNet34 [16] pre-trained on ImageNet [7] as their backbone.

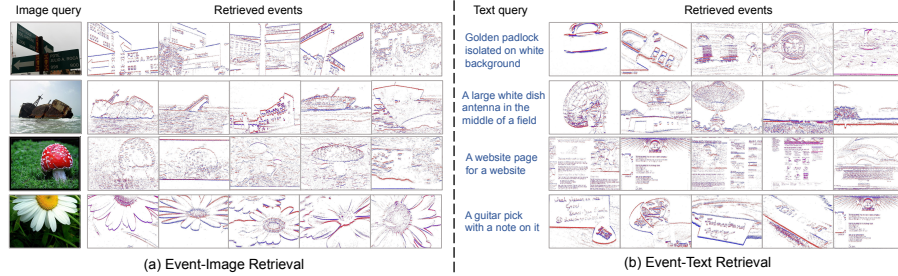


Fig. 3: Qualitative results of event-image retrieval and event-text retrieval.

Table 3: Quantitative results of event-image retrieval.

Method	N-ImageNet [20]						N-Caltech101 [35]					
	Event query			Image query			Event query			Image query		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
EventCLIP [45]	3.61	7.67	10.74	11.33	20.29	25.74	9.53	24.46	33.54	11.96	23.77	28.47
E2VID-CLIP	2.49	4.38	5.43	5.79	11.40	14.62	23.24	50.02	63.69	<u>35.69</u>	<u>60.32</u>	<u>72.90</u>
CETA-Prompt-S	2.43	5.73	7.91	11.91	23.53	29.85	7.58	23.24	31.76	16.14	38.94	52.21
CETA-BLIP2-S	4.82	10.08	13.75	15.61	29.08	35.74	10.54	27.13	36.34	19.26	42.83	55.78
CEIA-S	<u>35.24</u>	<u>54.33</u>	<u>61.73</u>	<u>36.62</u>	<u>54.73</u>	<u>62.72</u>	<u>28.19</u>	<u>52.98</u>	<u>64.94</u>	33.63	60.12	71.68
CEIA-L	41.22	61.39	69.61	44.15	62.86	70.12	32.94	60.56	72.45	39.83	66.36	78.09

As shown in Tab. 2, in the extreme case of 1-shot, the performance of Sorted Time Surface and DiST drops significantly due to the serious lack of training data. In contrast, our CEIA can leverage CLIP’s outstanding robustness to quickly adapt to the new distribution, demonstrating a large margin performance boost. In terms of Acc1, CEIA-L outperforms Sorted Time Surface by 43.53% and 56.66% in terms of Acc1 on N-ImageNet and N-Caltech101 with 1-shot, respectively. Compared to EventCLIP, our CEIA-L achieves superior results on all datasets and all N-shot settings. This indicates that CEIA significantly enhances the transferability of knowledge from CLIP to event-based vision.

4.5 Event-Image Retrieval

Metrics. We measure the performance of event-image retrieval through computing recall at K (R@K) [25], which is defined as the fraction of queries for which the correct item is retrieved in the closest K points to the query.

Results. Tab. 3 shows that our CEIA-L consistently outperforms EventCLIP and E2VID-CLIP under all metrics across both datasets. Specifically, on N-ImageNet, CEIA-L surpasses EventCLIP and E2VID-CLIP by 37.61% and 38.73% in terms of R@1 for event queries, respectively. The underlying reason is that the contrastive loss we used is essential for multi-modal retrieval as it directly learns cross-modal similarity and alleviates the domain disparity of event and image data. When compared to CETA-Prompt and CETA-BLIP2, CEIA also holds overwhelming advantages because it directly aligns event-image data. For instance, CEIA-S outperforms CETA-Prompt-S and CETA-BLIP2-S by 32.81%

Table 4: Quantitative results of event-text retrieval on our built event-text dataset NIN-BLIP2-retrieval.

Method	Event query			Text query		
	R@1	R@5	R@10	R@1	R@5	R@10
EventCLIP [45]	15.27	34.19	42.47	22.87	43.59	53.35
E2VID-CLIP	5.95	14.72	19.91	12.11	23.03	27.95
CETA-Prompt-S	14.67	33.51	42.44	23.16	44.15	54.51
CETA-BLIP2-S	22.80	47.03	57.31	<u>29.03</u>	52.63	<u>62.55</u>
CEIA-S	<u>24.75</u>	<u>49.27</u>	<u>58.55</u>	28.95	<u>52.75</u>	62.27
CEIA-L	29.44	57.44	66.51	34.23	59.79	69.47

and 30.42% in terms of R@1 for event queries on N-ImageNet. Additionally, we visualize the results of the image query on N-ImageNet in Fig. 3, where the retrieved events have a very high degree of similarity to the input image query.

4.6 Event-Text Retrieval

Metrics. Similar to the event-image retrieval task, we reuse the recall at K (R@K) [25] to evaluate the performance of the event-text retrieval task.

Results. As shown in Tab. 4, we report the results on our built event-text dataset N-ImageNet-BLIP2. As can be seen, our CEIA-L outperforms EventCLIP and E2VID-CLIP by a large margin in both event query and text query. For example, CEIA-L achieves an 11.36% improvement in terms of R@1 for text query compared to EventCLIP. Although CETA-BLIP2-S achieves slightly better results than our CEIA-S for text query, it’s an unfair comparison as CETA-BLIP2-S employs the captions generated by BLIP2 for training, which have the same distribution as N-ImageNet-BLIP2. In addition, we qualitatively show the results of CEIA in Fig. 3. Even when the caption describes the relationship of multiple objects (the 4th row), CEIA is able to accurately retrieve the most correlated events.

4.7 Domain Adaptation

Setting. We conduct domain adaptation based on object recognition, which aims to validate the effectiveness of enhancing event-based understanding by transferring the knowledge of the frame-based vision. Specifically, We first train a classifier as the task network using labeled data from the image domain, and then directly transfer it to the event domain.

Results. As observed in Tab. 5, our CEIA-L consistently secures the top position on both N-ImageNet and N-Caltech101. Specifically, in terms of Acc1, CEIA-L outperforms E2VID-CLIP by 39.30% on N-ImageNet and by 4.05% on N-Caltech101. Moreover, compared to CETA-Prompt-S and CETA-BLIP2-S, our CEIA-S also exhibits its superiority, achieving significant increases for all metrics. These remarkable improvements demonstrate that CEIA effectively

Table 5: Quantitative results of domain adaptation.

Method	N-ImageNet [20]		N-Caltech101 [35]	
	Acc1	Acc5	Acc1	Acc5
EventCLIP [45]	21.92	43.24	69.22	80.46
E2VID-CLIP	10.86	23.88	<u>81.12</u>	<u>93.76</u>
CETA-Prompt-S	21.32	38.58	70.83	83.10
CETA-BLIP2-S	27.66	47.40	72.20	84.04
CEIA-S	<u>43.64</u>	<u>70.56</u>	79.38	91.13
CEIA-L	50.16	76.36	85.17	94.89

Table 6: Ablation study on the effects of LoRA on event-text retrieval and domain adaptation.

Method	Event-text retrieval						Domain adaptation			
	Event query			Image query			N-ImageNet [20]		N-Caltech101 [35]	
	R@1	R@5	R@10	R@1	R@5	R@10	Acc1	Acc5	Acc1	Acc5
w/o LoRA	28.99	55.75	64.95	31.23	57.95	67.79	48.24	74.02	83.82	94.46
w/ LoRA	29.44	57.44	66.51	34.23	59.79	69.47	50.16	76.36	85.17	94.89

aligns event and image data within the same embedding space. Consequently, CEIA facilitates a smoother transfer of the task network trained in the image domain to the event domain, thereby enhancing the performance of domain adaptation. We believe that these performance gains can be transferred to other tasks and unlock the virtually unlimited image-based datasets for event-based vision, which will be our future work.

5 Ablation Study

The Effectiveness of LoRA. We compare two methods of training the event encoder: full finetuning and LoRA-based finetuning [18]. From Tab. 6, we can observe that, LoRA-based finetuning consistently outperforms full finetuning across all metrics for event-text retrieval and domain adaptation tasks. These results demonstrate that LoRA can effectively preserve CLIP’s strong robustness and meanwhile avoid overfitting to the training datasets.

LoRA Configuration. In Tab. 7, we evaluate various LoRA [18] configurations as depicted in Fig. 2. “ r ” represents the low intrinsic dimension of rank decomposition matrices. “ α ” indicates the scaling degree applied to the outputs from the trainable weights, and “Weight Type” denotes which weight matrices in the event encoder are finetuned with LoRA. The experimental results demonstrate that adapting only W_q and W_v with a very small r has already achieved competitive performance. Further increasing r or adjusting LoRA with more weights does not lead to significant improvements. Additionally, we set α as twice r to scale up the output from trainable weights, thereby further speeding up training.

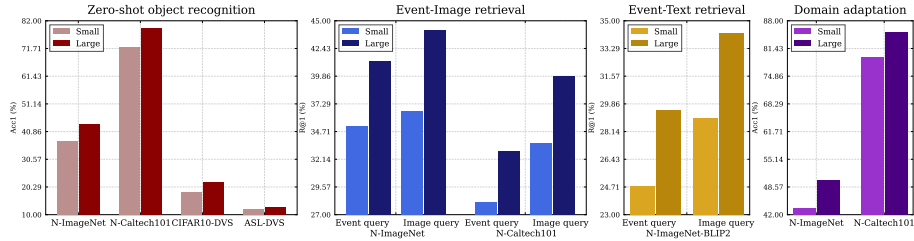


Fig. 4: Comparison of training CEIA with different scale data.

Table 7: Ablation study on LoRA configuration on N-ImageNet-S.

Rank r	16	32	64	16	16
LoRA α	16	32	64	32	32
Weight	W_q	W_q	W_q	W_q	W_q, W_k
Type	W_v	W_v	W_v	W_v	W_v, W_o
Acc1	36.28	36.80	36.80	<u>37.25</u>	37.33

Table 8: Ablation study on event representations on N-ImageNet-S.

Representations	Acc1	Acc5
DiST [20]	33.34	57.64
Time Surface [24]	34.84	58.74
Voxel [56]	30.91	54.24
Gray [45]	<u>35.94</u>	<u>60.31</u>
R-B [45]	37.25	61.60

Data Scalable Capability. As shown in Fig. 4, we can see that CEIA-L, which is trained on the larger-scale event-image pairs, achieves significantly better performance than CEIA-S across all benchmarks. This indicates that, leveraging more training data, CEIA can exhibit the flexibility to boost performance, ensuring its scalable capability. Therefore, larger-scale event-image pretraining is an exciting direction for future work.

Event Representations. The results in Tab. 8 show the ablation results of different event representations. Compared with the commonly-used DiST [20], Time Surface [24], Voxel [56], and Gray [45], the red-blue color map (referred to as R-B) [45] leads to the best recognition accuracy. We speculate that these worse results may be due to larger differences between these representations and natural images used by CLIP.

6 Conclusion

In this paper, we propose CEIA, an effective framework to adapt CLIP to event data. We provide a novel perspective of focusing on learning to align event and image data as an alternative, thus overcoming the challenge posed by the shortage of event-text datasets. We thoroughly evaluate CEIA on four applications: object recognition, event-image retrieval, event-text retrieval, and domain adaptation. The state-of-the-art results show that CEIA not only enhances open-world understanding but also opens the door to more event-based multi-modal understanding tasks. Furthermore, CEIA’s significant scalability under abundant event-image pairs also opens up the possibility to introduce the first event-based Large Vision Model, which will be our future work.

References

1. Alzugaray, I., Chli, M.: Ace: An efficient asynchronous corner tracker for event cameras. In: 2018 International Conference on 3D Vision (3DV). pp. 653–661. IEEE (2018)
2. Berner, R., Brandli, C., Yang, M., Liu, S.C., Delbruck, T.: A 240×180 10mw 12us latency sparse-output vision sensor for mobile applications. In: 2013 Symposium on VLSI Circuits. pp. C186–C187. IEEE (2013)
3. Bi, Y., Chadha, A., Abbas, A., Boursoulatz, E., Andreopoulos, Y.: Graph-based spatio-temporal feature learning for neuromorphic vision sensing. *IEEE Transactions on Image Processing* **29**, 9084–9098 (2020)
4. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A 240×180 130 db $3 \mu\text{s}$ latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits* **49**(10), 2333–2341 (2014)
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
6. Das Biswas, S., Kosta, A., Liyanagedera, C., Apolinario, M., Roy, K.: Halsie: Hybrid approach to learning segmentation by simultaneously exploiting image and event modalities. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5964–5974 (2024)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
9. Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H.R.: A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020* pp. 877–894 (2021)
10. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop. pp. 178–178. IEEE (2004)
11. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* **44**(1), 154–180 (2020)
12. Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D.: End-to-end learning of representations for asynchronous event-based data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5633–5643 (2019)
13. Glover, A., Vasco, V., Bartolozzi, C.: A controlled-delay event camera framework for on-line robotics. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 2178–2183. IEEE (2018)
14. Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M.: Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence* **43**(12), 4338–4364 (2020)
15. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Audioclip: Extending clip to image, text and audio. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 976–980. IEEE (2022)

16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. Hegde, D., Valanarasu, J.M.J., Patel, V.: Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2028–2038 (2023)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
19. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
20. Kim, J., Bae, J., Park, G., Zhang, D., Kim, Y.M.: N-imagenet: Towards robust, fine-grained object recognition with event cameras. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2146–2156 (2021)
21. Kim, Y., Chough, J., Panda, P.: Beyond classification: Directly training spiking neural networks for semantic segmentation. *Neuromorphic Computing and Engineering* **2**(4), 044015 (2022)
22. Klenk, S., Bonello, D., Koestler, L., Araslanov, N., Cremers, D.: Masked event modeling: Self-supervised pretraining for event cameras. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2378–2388 (2024)
23. Krizhevsky, A.: Learning multiple layers of features from tiny images. Master’s thesis, University of Tront (2009)
24. Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.B.: Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence* **39**(7), 1346–1359 (2016)
25. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European conference on computer vision (ECCV). pp. 201–216 (2018)
26. Li, H., Liu, H., Ji, X., Li, G., Shi, L.: Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience* **11**, 309 (2017)
27. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
28. Li, N., Bhat, A., Raychowdhury, A.: E-track: Eye tracking with event camera for extended reality (xr) applications. In: 2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS). pp. 1–5. IEEE (2023)
29. Lin, S., Zhang, J., Pan, J., Jiang, Z., Zou, D., Wang, Y., Chen, J., Ren, J.: Learning event-driven video deblurring and interpolation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. pp. 695–710. Springer (2020)
30. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
32. Messikommer, N., Gehrig, D., Gehrig, M., Scaramuzza, D.: Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters* **7**(2), 3515–3522 (2022)

33. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition. In: European Conference on Computer Vision. pp. 1–18. Springer (2022)
34. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
35. Orchard, G., Jayawant, A., Cohen, G.K., Thakor, N.: Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience* **9**, 437 (2015)
36. Paikin, G., Ater, Y., Shaul, R., Soloveichik, E.: Efi-net: Video frame interpolation from fusion of events and frames. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1291–1301 (2021)
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
38. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Dense-clip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18082–18091 (2022)
39. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3857–3866 (2019)
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
41. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
42. Sun, Z., Messikommer, N., Gehrig, D., Scaramuzza, D.: Ess: Learning event-based semantic segmentation from still images. In: European Conference on Computer Vision. pp. 341–357. Springer (2022)
43. Tulyakov, S., Bochicchio, A., Gehrig, D., Georgoulis, S., Li, Y., Scaramuzza, D.: Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17755–17764 (2022)
44. Weng, W., Zhang, Y., Xiong, Z.: Event-based blurry frame interpolation under blind exposure. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1588–1598 (2023)
45. Wu, Z., Liu, X., Gilitschenski, I.: Eventclip: Adapting clip for event-based object recognition. arXiv preprint arXiv:2306.06354 (2023)
46. Xiao, Z., Weng, W., Zhang, Y., Xiong, Z.: Eva²: Event-assisted video frame interpolation via cross-modal alignment and aggregation. *IEEE Transactions on Computational Imaging* **8**, 1145–1158 (2022)
47. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1179–1189 (2023)
48. Yang, Y., Pan, L., Liu, L.: Event camera data pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10699–10709 (2023)
49. Yu, Z., Zhang, Y., Liu, D., Zou, D., Chen, X., Liu, Y., Ren, J.S.: Training weakly supervised video frame interpolation with events. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14589–14598 (2021)

- 50. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
- 51. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8552–8562 (2022)
- 52. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
- 53. Zhang, X., Yu, L.: Unifying motion deblurring and frame interpolation with events. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17765–17774 (2022)
- 54. Zhou, C., Teng, M., Han, J., Liang, J., Xu, C., Cao, G., Shi, B.: Deblurring low-light images with events. *International Journal of Computer Vision* **131**(5), 1284–1298 (2023)
- 55. Zhou, J., Zheng, X., Lyu, Y., Wang, L.: E-clip: Towards label-efficient event-based open-world understanding by clip. arXiv preprint arXiv:2308.03135 (2023)
- 56. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–997 (2019)