# Label-Free Event-based Object Recognition via Joint Learning with Image Reconstruction from Events

Hoonhee Cho*, Hyeonseong Kim*, Yujeong Chae, and Kuk-Jin Yoon
Korea Advanced Institute of Science and Technology
{gnsgnsgml,brian617,yujeong,kjyoon}@kaist.ac.kr

## Abstract

*Recognizing objects from sparse and noisy events becomes extremely difficult when paired images and category labels do not exist. In this paper, we study label-free event-based object recognition where category labels and paired images are not available. To this end, we propose a joint formulation of object recognition and image reconstruction in a complementary manner. Our method first reconstructs images from events and performs object recognition through Contrastive Language-Image Pre-training (CLIP), enabling better recognition through a rich context of images. Since the category information is essential in reconstructing images, we propose category-guided attraction loss and category-agnostic repulsion loss to bridge the textual features of predicted categories and the visual features of reconstructed images using CLIP. Moreover, we introduce a reliable data sampling strategy and local-global reconstruction consistency to boost joint learning of two tasks. To enhance the accuracy of prediction and quality of reconstruction, we also propose a prototype-based approach using unpaired images. Extensive experiments demonstrate the superiority of our method and its extensibility for zero-shot object recognition. Our project code is available at* https://github.com/Chohoonhee/Ev-LaFOR.

## 1. Introduction

Event cameras are neuromorphic vision sensors that asynchronously perceive per-pixel brightness changes with high temporal resolution. They show advantages over conventional frame-based cameras, such as high dynamic range (HDR) and low latency. However, despite the superiority of cameras, algorithms using events are still in their infancy. Existing event-based object recognition methods [1, 4, 22, 24, 25, 26, 30, 32, 34, 36, 40, 53] have been shown that object recognition is a crucial task for event feature learning [12, 29, 38] and endeavors have successfully
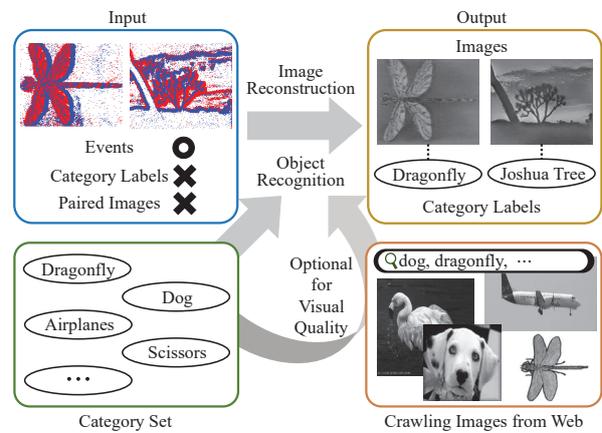


Figure 1: We tackle label-free event-based object recognition where category labels and paired images are not available. Our approach can simultaneously perform object recognition and image reconstruction through proposed joint learning framework. Optionally, our method can leverage unpaired images to enhance object recognition performance and reconstructed image quality.

classified the object with events. However, they rely on supervision with category labels to achieve high performance. Furthermore, the sparsity of events poses a challenge for users to label them accurately when there are no paired images. Several unsupervised methods [25, 36] have been proposed for object recognition, however, their performances remain subpar. Hence, there is a natural need for robust and effective label-free event-based object recognition methods.

In this paper, we study label-free event-based object recognition where category labels and paired images are not available as in Fig. 1. To tackle this task, we focus on two following facts. Firstly, as demonstrated in [45, 60], performing object recognition on reconstructed images with spatially dense information and rich context shows better performance than solely using events. Secondly, during reconstructing images from events, taking the categories into

---

*The first two authors contributed equally. In alphabetical order.

account can improve the quality of the reconstructed images. From these observations, we propose a joint learning framework of two closely-linked tasks: event-based object recognition and event-to-image reconstruction.

In our joint learning framework, we first reconstruct images from events and perform object recognition on the reconstructed images using Contrastive Language-Image Pre-training (CLIP) [42]. CLIP is an image-text embedding model that has been trained on a large number of image and language pairs sourced from the web. We utilize the textual features encoded from CLIP since they are well-aligned with visual features, which is advantageous for our tasks that lack labels and images. To this end, we propose the category-guided attraction loss that aligns the visual features from reconstructed images close to the texture features of the predicted category from CLIP.

Although the above framework enables the joint learning of event-based object recognition and image reconstruction, using the predicted categories brings two issues. (1) Collapsing to the trivial solution, resulting in all outputs being assigned to a single category and (2) Interrupting the learning process due to unreliable (wrong) predictions. To alleviate these problems, we propose (1) category-agnostic repulsion loss to increase the distances between visual features for preventing collapse and (2) a reliable data sampling (RDS) method to reduce the effects of unreliable predictions. For RDS, we devise two reliable indicators named posterior probability indicator (PPI) and temporally reversed consistency indicator (TRCI). The PPI selects reliable samples based on the probability of prediction. For additional sampling, TRCI utilizes the characteristics of events that temporally reversed ones should be categorized into the same category as the original events. This reliable data sampling enables stable learning even with the predicted categories, leading to a substantial performance improvement in object recognition. Furthermore, to restore the local details, we introduce the local-global reconstruction consistency. Finally, to enhance the accuracy of recognition and visual quality of reconstructed images, we expand our method by employing the existing unpaired images. Extensive experiments demonstrate that our method shows outstanding results on the subset of the N-ImageNet dataset over unsupervised and image reconstruction-based approaches and even surpasses the supervised methods on the N-Caltech101 dataset without using labels and images. In addition, we show that our framework can be extended to event-based zero-shot object recognition.

Our work presents four main contributions: (I) We propose a novel joint learning framework for event-based object recognition and image reconstruction, without requiring paired images and labels. (II) We introduce a reliable data sampling strategy and local-global reconstruction consistency that enhances label-free joint training. (III) We further introduce employing unpaired images in our framework for boosting performance. (IV) We conduct extensive experiments of our method along with zero-shot object recognition and superset categories learning, and demonstrate our method's superior performance and effectiveness.

## 2. Related Work

**Event-based Object Recognition.** Concurrent with the success of event-based low-level tasks [9, 10, 11, 23, 52, 54, 63], there has been a surge of active research focused on object recognition [3, 7, 19, 22, 27, 28, 31, 35, 43, 51, 53, 56], which is considered as a fundamental task. Considering the unique features of event data, numerous works [1, 4, 22, 24, 25, 26, 30, 32, 34, 36, 40, 53] have presented models and representations that can effectively extract significant patterns. However, these approaches require labels for each event data for supervised learning, and since event data cannot be crawled from the web like images, the labeling cost of event data is extremely high. On the other hand, our approach enables the learning of event-based object recognition and image reconstruction simultaneously, without requiring category labels and paired images.

**Event-based Image Reconstruction** is a well-known topic in the field of event-based vision [2, 13, 17, 20, 21, 46, 49, 57]. Recent researches have been utilizing deep learning to generate realistic outcomes [44, 45, 50, 62]. However, these supervised approaches require the paired images of each event stream to be accurately aligned and synchronized in terms of pixels and time. Another line of approach, such as Wang *et al*. [59] and Pini *et al*. [41], utilized generative adversarial networks (GANs) to reconstruct intensity with unpaired images that are not related to the events of the scene. However, GANs are prone to be sensitive to various parameters, and when images that are significantly irrelevant to objects are utilized, the quality of the reconstructed images from events is notably low. An alternative solution is a self-supervised method [37] that utilizes photometric consistency, but this approach requires a continuous sequence that spans an adequately long period. Differently, our framework can reconstruct the intensity of event data without images, very long event sequence and GANs. We only exploit the category prompts of a bundle of event data, not the class label of each event data.

**Label-free Learning** aims to solve the end-task only with unlabeled data, and has been actively explored [8, 58]. HOTS [25] proposed a category-free method for recognizing objects based on events and SSL-E2VID [37] proposed a self-supervised learning method for reconstructing images from only events. However, their performances fall short of supervised learning methods with significant performance gaps. Inspired by the success of using the pre-trained model [5, 6, 39, 48, 55], we adopt CLIP [42] for label-free event-based object recognition. We propose a
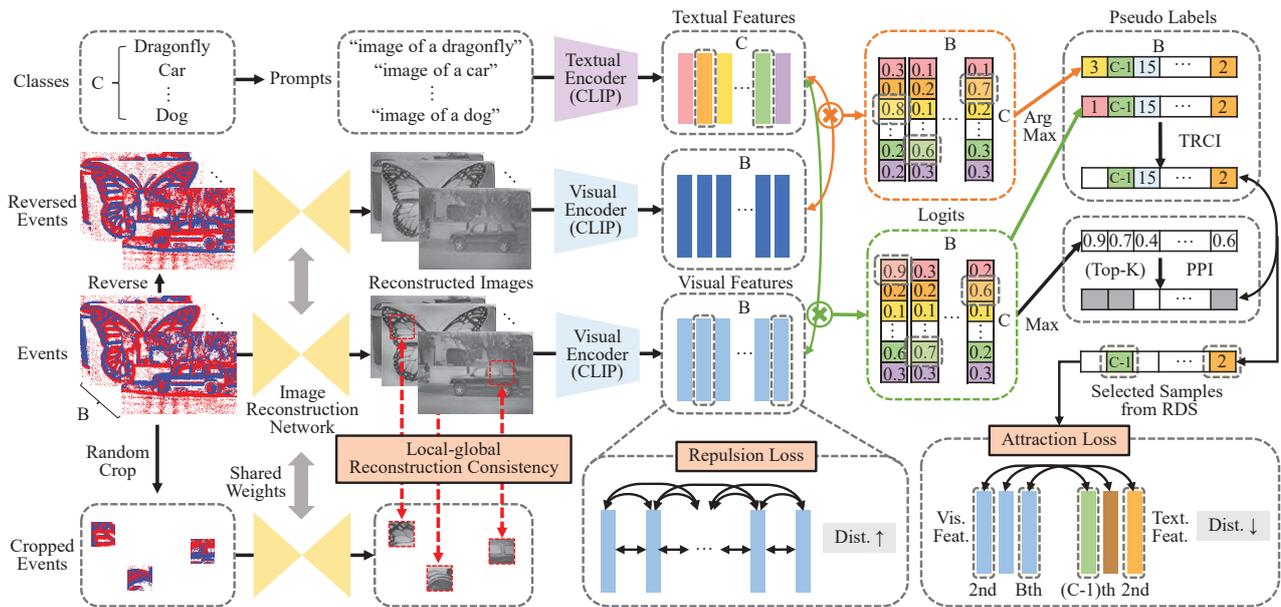
Figure 2: Overall joint learning framework of event-based object recognition with event-to-image reconstruction. Our approach involves CLIP's textual and visual encoders that derive textual and visual features, respectively, from given class sets and events. Textual features are obtained from category prompts, while visual features are obtained from reconstructed images. Then, logits and corresponding pseudo labels are generated by computing the cosine similarity between visual and textual features. Lastly, our framework employs pseudo labels to facilitate joint learning using category-agnostic repulsion loss and category-guided attraction loss. To mitigate the impact of unreliable pseudo labels and boost performance, we introduce a reliable data sampling strategy and local-global reconstruction consistency.

novel joint learning framework that uses visual and textual knowledge to achieve superior performance in event-to-image reconstruction and event-based object recognition without category labels and paired images. To the best of our knowledge, we are the first to incorporate textual knowledge into event vision research, particularly in the context of label-free tasks.

## 3. Method

In Sec. 3.1, we first introduce our joint formulation of event-to-image reconstruction and object recognition for label-efficient learning. In Sec. 3.2, we describe our reliable data sampling (RDS) strategy for boosting joint learning performance by filtering the unreliable data samples during training. Sec. 3.3 introduces the local-global reconstruction consistency for spatially regularizing the reconstructed images, which ensures capturing the local details. In Sec. 3.4, we propose to further employ the unpaired images for better recognition performance and reconstructed image quality.

### 3.1. Joint Formulation

We propose a joint formulation of event-to-image reconstruction and object recognition to perform event-based (zero-shot) object recognition without using labels and paired images of event data. Our joint formulation is based on the fact that 1) image reconstruction can be helpful in

predicting categories from events, and 2) category information is important in reconstructing images from events. Therefore, we propose 1) the process of performing event-to-image reconstruction and CLIP-driven object recognition from the reconstructed images and 2) category-guided attraction loss and category-agnostic repulsion loss for learning this process without labels and paired images. The overall framework is shown in Fig. 2. Note that our joint learning framework can reconstruct intensity images from events without using any images.

**Event-to-Image Reconstruction.** For the joint formulation, we first reconstruct the intensity image $\mathcal{I}$ from the event stream $\mathcal{E} = (u_i, t_i, p_i)_{i=1}^N$, where $u_i$, $t_i$, and $p_i$ denote triggered pixel location, time, and polarity of each event, and $N$ is the temporal length of the event stream. We convert the event stream into event spiking tensor (EST) [16], which is then fed into the image reconstruction network $\mathcal{G}$ to obtain the intensity image. The image reconstruction process is formulated as follows:

$$\mathcal{I} = \mathcal{G}(\text{EST}(\mathcal{E})). \tag{1}$$

**CLIP-driven Object Recognition.** We perform object recognition on top of the reconstructed images instead of raw event data. The advantage of using reconstructed images rather than using event data directly for object recognition is that better performance can be achieved by using

a network trained on rich image data [45, 60]. In particular, we use CLIP [42] for object recognition, which also enables zero-shot object recognition. For an event dataset of $C$ classes, the category prompts are constructed by placing each category name into a pre-defined template, "image of a [CLASS]." Then, CLIP's textual encoder encodes category prompts to obtain the $D$-dimensional textual features $F = (f_i)_{i=1}^C \in \mathbb{R}^{C \times D}$. Meantime, the reconstructed image $\mathcal{I}$ is encoded by CLIP's visual encoder to a visual feature $v \in \mathbb{R}^{1 \times D}$ and the class probability $p \in \mathbb{R}^{1 \times C}$ for the object recognition are obtained as,

$$p_i = \text{softmax}_i(vF^\top), \tag{2}$$

where $p_i$ and $\text{softmax}_i(\cdot)$ denote the probability for category $i$ and the softmax function. By taking the category of maximum probability, we can predict the category $c$ for the event data $\mathcal{E}$ as,

$$c = \underset{i}{\arg\max}(p_i). \tag{3}$$

**Category-guided Attraction Loss.** The challenging part of our joint formulation is to train the reconstruction network $\mathcal{G}$ without using paired images of event data. Once we reconstruct the intensity images from the event stream (Eq. 1), we can perform object recognition using CLIP (Eq. 2). Motivated by the fact that category information is important in reconstructing images from events, we propose category-guided attraction loss to train the reconstruction network. The proposed category-guided attraction loss brings the visual feature $v$ encoded from the reconstructed image $\mathcal{I}$ closer to the textual feature of the corresponding category, guiding the model to reconstruct the image corresponding to the category. Since category labels are not available for the event data, we use the predicted categories as pseudo labels. For event data with a batch size of $B$, the category-guided attraction loss $\mathcal{L}_{att}$ is formulated as InfoNCE loss [33]:

$$\mathcal{L}_{att} = -\sum_{i=1}^B \log \frac{\exp(v_i f_{c_i}^\top)}{\sum_{j=1}^B \exp(v_i f_{c_j}^\top)}, \tag{4}$$

where $v_i$, $c_i$, and $f_j$ denote the visual feature and predicted category of the $i^{\text{th}}$ event data, and the textual feature of category $j$. Minimizing $\mathcal{L}_{att}$ enables the image reconstruction network to generate images of the corresponding category label, while object recognition benefits from the reconstructed images.

**Category-agnostic Repulsion Loss.** Although $\mathcal{L}_{att}$ enables joint learning, the use of pseudo labels may cause collapsing, resulting in all outputs being assigned to a single category. Therefore, we introduce a category-agnostic repulsion loss to prevent collapsing by increasing the distances between visual features. To be specific, the category-agnostic repulsion loss $\mathcal{L}_{rep}$ is applied to the set of visual features as follows:

$$\mathcal{L}_{rep} = \sum_{i=1}^B \log(1 + \sum_{j=1, j \neq i}^B \exp(v_i v_j^\top)). \tag{5}$$

With $\mathcal{L}_{att}$ and $\mathcal{L}_{rep}$, we can successfully formulate our joint learning framework, enabling simultaneous event-to-image reconstruction and (zero-shot) object recognition.

### 3.2. Reliable Data Sampling

As $\mathcal{L}_{att}$ relies on pseudo labels, the reliability of predictions may affect the overall training procedure and performance. To reduce the effects of unreliable predictions, we devise two reliability indicators named posterior probability indicator (PPI) and temporally reversed consistency indicator (TRCI). Based on the resultant reliability of predictions, we sample data with reliable predictions for $\mathcal{L}_{att}$. The remaining unreliable data is excluded from calculating $\mathcal{L}_{att}$ and used only for $\mathcal{L}_{rep}$.

**Posterior Probability Indicator.** Given the reconstructed image $\mathcal{I}$ and the predicted category $c$, the posterior probability $p_c = \mathbf{p}(c|\mathcal{I})$ indicates the model confidence for category $c$. Thus, we use the posterior probability as an indicator of the reliability of the prediction. In particular, we select $K$ data samples corresponding to the top-$K$ posterior probabilities among $B$ data samples in the mini-batch. The set of selected sample indices from PPI, $S_{\text{PPI}}$ is defined as,

$$S_{\text{PPI}} = [\underset{i}{\arg\text{sort}}(\mathbf{p}(c_i|\mathcal{I}_i))]_K, \tag{6}$$

where $\mathcal{I}_i$ and $[\cdot]_K$ denote the reconstructed image of the $i^{\text{th}}$ event data in the mini-batch and the set of the first $K$ items. The $\arg\text{sort}(\cdot)$ function returns indices that would sort an input in descending order.

**Temporally Reversed Consistency Indicator.** Considering the temporal aspect of event data, we propose an additional reliability indicator based on the consistency between predictions from the original events and the temporally reversed events. From the fact that the temporally reversed events $\mathcal{E}^R = (u_i, \max_j(t_j) - t_i, p_i)_{i=N}^1$ should be categorized into the same category as the original events $\mathcal{E}$, we presume that the consistent model prediction would indicate the reliability of the data sample. To be specific, the set of selected sample indices from TRCI, $S_{\text{TRCI}}$ is defined as follows:

$$S_{\text{TRCI}} = \{i | c_i = c_i^R, \ i = 1, 2, \cdots, B\}, \tag{7}$$

where $c_i^R$ denotes a predicted category of $i^{\text{th}}$ reversed event data $\mathcal{E}_i^R$ in the mini-batch.

Our final set of selected sample indices $S_{\text{RDS}}$ for reliable data sampling (RDS) is the intersection of $S_{\text{PPI}}$ and $S_{\text{TRCI}}$:

$$S_{\text{RDS}} = S_{\text{PPI}} \cap S_{\text{TRCI}}. \tag{8}$$

Then, the modified attraction loss with RDS is as follows:

$$\mathcal{L}_{att} = -\sum_{i \in S_{\text{RDS}}} \log \frac{\exp(v_i f_{c_i}^\top)}{\sum_{j \in S_{\text{RDS}}} \exp(v_i f_{c_j}^\top)}, \tag{9}$$

## 3.3. Local-global Reconstruction Consistency

Along with $\mathcal{L}_{att}$ and $\mathcal{L}_{rep}$, we further introduce the local-global reconstruction consistency for spatially regularizing the reconstructed images in a self-supervised manner. With random crop function $\mathcal{H}$, we first crop the EST of original events $\text{EST}(\mathcal{E})$ to obtain the cropped tensor $\mathcal{H}(\text{EST}(\mathcal{E}))$ for the local input. Then, the local-global reconstruction consistency loss $L_{con}$ is defined as,

$$\mathcal{L}_{con} = \|\mathcal{G}(\mathcal{H}(\text{EST}(\mathcal{E}))) - \mathcal{H}(\mathcal{G}(\text{EST}(\mathcal{E})))\|_1. \quad (10)$$

Compared to $\mathcal{L}_{att}$ and $\mathcal{L}_{rep}$ that work on visual features, $L_{con}$ is applied directly to the reconstructed images and considered as a spatial regularization for the reconstruction network $\mathcal{G}$. In addition, matching the reconstructed images from local and global inputs enables the network to capture the local details, improving the reconstruction quality.

The total loss function for the joint learning framework is as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{att} + \lambda_2 \mathcal{L}_{rep} + \lambda_3 \mathcal{L}_{con}, \quad (11)$$

where $\lambda_1$, $\lambda_2$, and $\lambda_3$ are weights for $\mathcal{L}_{att}$, $\mathcal{L}_{rep}$, and $\mathcal{L}_{con}$.

## 3.4. Employing Unpaired Images

In this section, we expand the use case of our method by introducing how to employ the existing unpaired images. Here, we assume that the category of each image is known, which is the common case for web crawling. Although our framework is capable of event-to-image reconstruction and object recognition without using any images and labels, utilizing existing images can possibly improve the image reconstruction quality and consequently improve object recognition performance. As shown in Fig. 3, we replace textual features in $\mathcal{L}_{att}$ with the prototype features generated from unpaired images. For each category $c$, we first encode all images of category $c$ and apply a clustering algorithm to obtain $L$ clusters in total. Then, we aggregate the features of each cluster by averaging them, resulting in $L$ prototype features $W^c = (w_i^c)_{i=1}^L \in \mathbb{R}^{L \times D}$.

For data sample index $i \in S_{\text{RDS}}$ with corresponding visual feature $v_i$ and the predicted category $c_i$, we replace the textual features $f_{c_i}$ in Eq. 9 to the closest prototype feature $w_{l_i}^{c_i}$, where the cluster index $l_i$ is calculated as,

$$l_i = \underset{j}{\arg\max}(v_i(w_j^{c_i})^\top). \quad (12)$$

Our modification of $\mathcal{L}_{att}$ to employ unpaired images is then defined as follows:

$$\mathcal{L}_{att} = - \sum_{i \in S_{\text{RDS}}} \log \frac{\exp(v_i(w_{l_i}^{c_i})^\top)}{\sum_{j \in S_{\text{RDS}}} \exp(v_i(w_{l_j}^{c_j})^\top)}, \quad (13)$$
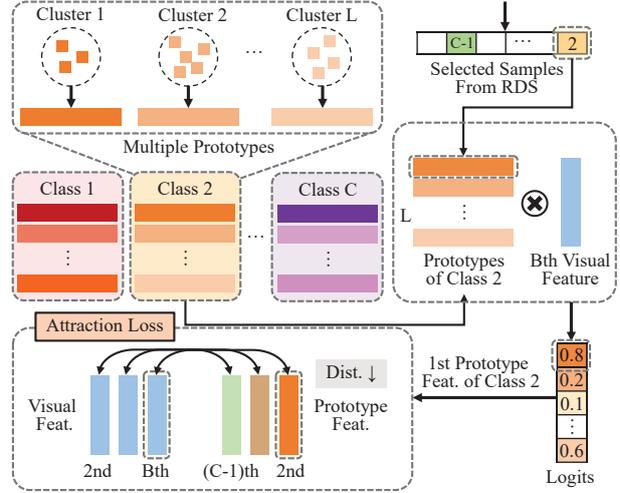


Figure 3: Proposed method employing unpaired images in the category-guided attraction loss. Unpaired images are further utilized to enhance object recognition performance and reconstructed image quality.

## 4. Datasets

The experiments are conducted on two widely used event-based object recognition datasets, N-Caltech101 [35] and N-ImageNet [22], which are constructed by filming Caltech101 [15] and ImageNet [14] datasets on LCD monitor with event camera. Since N-Caltech101 and N-ImageNet datasets do not provide APS frames, the image in the original dataset and corresponding event sequence have different resolutions and are not aligned in pixel-level. To create a visual prototype in Sec. 3.4, we scrap images from the original dataset, mimicking the web crawling. To avoid including paired images of events in the train set, we randomly split the event train set in half for each class. One half is used to train the model, while the corresponding images from the other half are used to create the visual prototype.

**N-Caltech101** dataset consists of 8,246 event stream data belonging to 101 different object classes. Each object is captured with $320 \times 245$ resolution (average) for 300 milliseconds. Existing works [45, 60] utilized this dataset to evaluate the performance of image reconstruction and event-based object recognition.

**N-ImageNet** is recently released large-scale dataset, containing 1,781,167 event stream data of 1,000 object classes. Compared to N-Caltech101, it has a larger resolution of $640 \times 480$ and a very short capturing time with $5 \times 10^{-2}$ milliseconds. Therefore, each event stream contains significantly sparse edge information. N-ImageNet is too challenging for unsupervised studies without ground truth, so we utilize a randomly sampled subset of 100 classes called N-ImageNet (Mini). Details about the split of N-ImageNet (Mini) are provided in the *supplementary material*.

| Methods | Image Pair | Class Label | N-Caltech101 | | | N-ImageNet (Mini) | | |
|---|:---:|:---:|:---:|:---:|:---:|:---:|:---:|:---:|
| | | | Accuracy ↑ | FID ↓ | IS ↑ | Accuracy ↑ | FID ↓ | IS ↑ |
| **Supervised Learning** | | | | | | | | |
| Histogram [29] | | ✓ | *63.8* | - | - | *49.8* | - | - |
| Ev-gait [61] | | ✓ | *61.9* | - | - | *51.3* | - | - |
| EST [16] | | ✓ | *81.7** | - | - | *52.5* | - | - |
| DiST [22] | | ✓ | *73.8* | - | - | *57.9* | - | - |
| **Unsupervised Learning** | | | | | | | | |
| HOTS [25] | | | 21.0* | - | - | 1.6 | - | - |
| **Image Reconstruction + CLIP** | | | | | | | | |
| E2VID [44] (Pre-train) | | | 64.0 | 83.90 | 12.81 | 13.3 | 147.14 | 7.82 |
| E2VID [44] (Fine-tune) | | | 59.8 | 85.31 | 12.91 | 13.4 | 144.57 | 7.48 |
| E2VID [44] (Fine-tune) | ✓ | | 59.8 | 86.66 | 12.61 | 13.5 | <u>144.03</u> | 7.40 |
| E2VID [44] (Fine-tune) | | ✓ | 60.1 | 85.50 | 12.85 | 12.7 | 145.79 | 7.33 |
| E2VID [44] (Scratch) | ✓ | | 9.4 | 302.48 | 2.60 | 1.0 | 299.70 | 1.15 |
| SSL-E2VID [37] (Pre-train) | | | 28.2 | 281.09 | 2.36 | 1.6 | 324.43 | 2.08 |
| SSL-E2VID [37] (Scratch) | | | 30.5 | 228.30 | 4.37 | 1.7 | 250.55 | 3.00 |
| Wang *et al*. [59] (Scratch) | | | 42.7 | 160.34 | 6.83 | 10.1 | 231.94 | 5.00 |
| Wang *et al*. [59] (Scratch) | ✓ | | 43.5 | 72.67 | 4.50 | 19.0 | 180.76 | 4.89 |
| Wang *et al*. [59] (Scratch) | | ✓ | 39.7 | 179.79 | 4.18 | 13.2 | 216.37 | 4.70 |
| Ours (Text Prompt) | | | <u>82.46</u> | <u>62.29</u> | <u>14.81</u> | 30.16 | 150.58 | <u>7.92</u> |
| Ours (Visual Prototype) | | | **82.61** | **54.19** | **17.59** | **31.28** | **135.95** | **9.74** |

Table 1: The results of event-based object recognition and event-to-image reconstruction. The **bold** and <u>underline</u> denote the best and the second-best performance except for supervised methods, respectively. * denotes the values taken from the original paper and *italic* denotes the performance of supervised methods.

## 5. Experiments

We compare our methods utilizing text prompts or visual prototypes with other existing methods. We adopt four representations (Histogram [29], Ev-gait [61], EST [16], and DiST [22]) for supervised methods and an unsupervised method (HOTS [25]). Additionally, three image reconstruction networks, E2VID [44], SSL-E2VID [37] and Wang *et al*. [59], are utilized with CLIP. For a fair comparison with image reconstruction networks, we record the performance of three different models for each image reconstruction network: the official pre-trained model, fine-tuned model, and a model trained from scratch. We evaluate these models under three conditions: paired images, images belonging to the same class, and randomly selected images. Some cases are overlooked when the pre-trained model is unavailable or the training setup does not yield meaningful results. For implementation details, see the *supplementary material*.

### 5.1. Label-free Object Recognition

**Text Prompt.** The quantitative results are presented in Table 1. The proposed method outperforms supervised object recognition methods on the N-Caltech101 dataset, despite not using category labels. Also, our method surpasses existing unsupervised methods and image reconstruction methods with CLIP by a large margin under all

metrics. Specifically, ours surpasses the best image reconstruction with CLIP method, E2VID (Pre-train), by around 28% increase in accuracy even though it utilizes the image while pre-training. On N-ImageNet, supervised methods record higher accuracy than ours since the dataset is challenging due to the sparse events captured within a short period. However, ours shows a significant improvement in all metrics compared to other unsupervised and image reconstruction-based methods. Our approach demonstrates the ability to perform object recognition and image reconstruction from event data, even when trained solely on textual features without any image data.

**Visual Prototype.** Besides text prompts, our framework can employ visual prototypes generated from scrapped images. By incorporating visual information, our visual prototype-based approach exceeds the performance of relying solely on textual prompts and achieves the best performance among unsupervised and image reconstruction-based methods under all metrics across both datasets. This demonstrates the superiority of the visual prototypes.

### 5.2. Image-free Image Reconstruction

We also compare the reconstruction results of our methods with other event-to-image reconstruction networks in Table 1. Note that, as mentioned in Sec. 4, the resolution and pixel positions of events and corresponding orig-

| Events | Wang *et. al.* | E2VID | E2VID* | Ours (Prompt) | Ours (Proto.) | Ours (Proto.)* | Image |
|--------|------|-------|--------|---------------|---------------|----------------|-------|
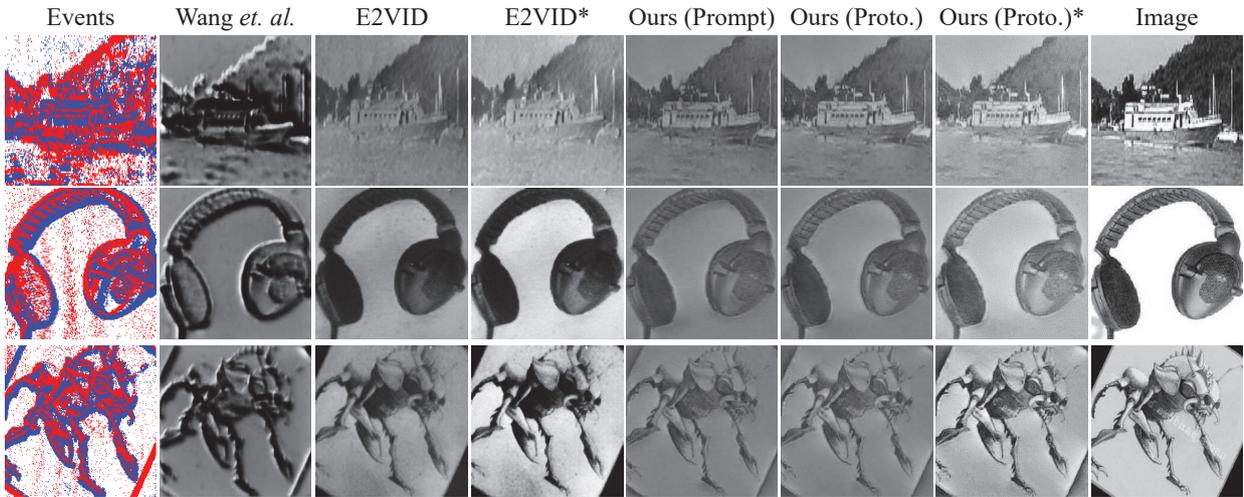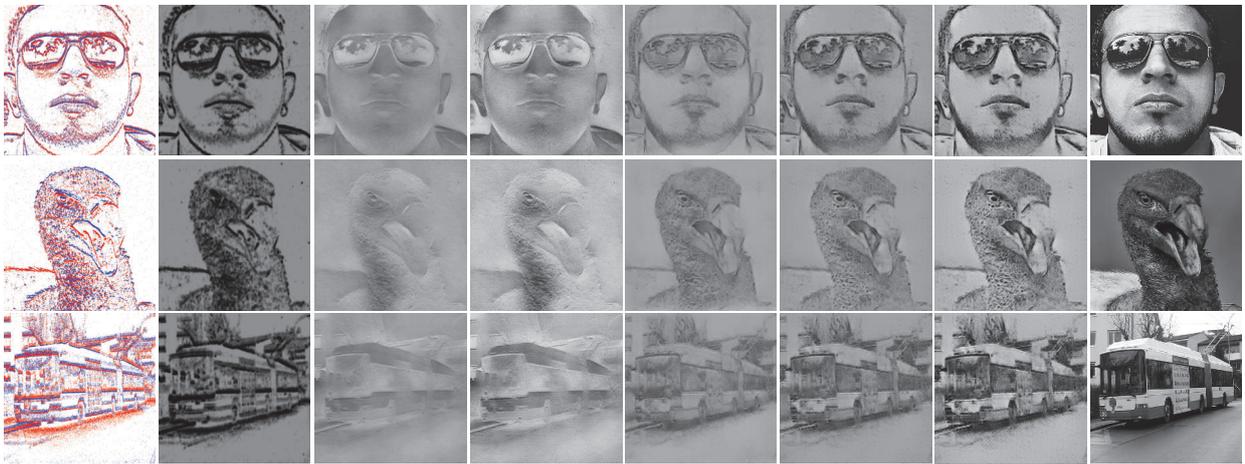


(a) N-Caltech101



(b) N-ImageNet

Figure 4: Visual results of event-to-image reconstruction network on (a) N-Caltech101 and (b) N-ImageNet datasets. Ours shows a higher quality of reconstructed images with fine-grained details than other methods (Wang *et al*. [59] and pre-trained E2VID [44]). * denotes the images with brightness and contrast adjustment.

inal images are not aligned. Therefore, pixel-wise metrics (*e.g.*, L1, PSNR) cannot be used for comparison. Instead, to assess the visual quality of the reconstructed images, we utilize two metrics: inception score (IS) [47] and Fréchet inception distance (FID) [18].

Despite not utilizing any images during training, our text prompt-based approach surpasses all other image reconstruction techniques in terms of FID and IS metrics on the N-Caltech101 dataset. However, in the case of N-ImageNet, our text prompt approach is the strongest by an IS indicator but falls behind pre-trained E2VID in FID. Since ours is trained from events of short periods without paired images, it becomes challenging to learn the distribution of real images. Instead, utilizing the visual prototype results in the highest FID and IS indicators in both datasets. This implies

that it can acquire the ability to reconstruct images without requiring paired images that correspond to events.

We also present the qualitative results in Fig. 4. Wang *et al*. [59] utilize the paired image in training but the method lacks the ability to effectively display intensity contrast, resulting in the loss of many details. While pre-trained E2VID shows better image quality over [59], it still struggles to preserve fine details as it operates in a distinct training environment. This is particularly apparent in the N-ImageNet dataset, where E2VID generates results that deviate significantly from the original image. Conversely, our framework is able to recover intensity effectively and preserve details in reconstructed images, even without access to paired images. Our results closely resemble real images, even in the challenging N-ImageNet dataset.

| Methods | Image Pair | Class Label | N-Caltech101 | | | N-ImageNet (Mini) | | |
|---|---|---|---|---|---|---|---|---|
| | | | Accuracy ↑ | FID ↓ | IS ↑ | Accuracy ↑ | FID ↓ | IS ↑ |
| **Image Reconstruction + CLIP** | | | | | | | | |
| E2VID [44] (Pre-train) | | | 58.4 | 194.26 | 6.47 | 31.7 | 203.05 | 6.64 |
| E2VID [44] (Fine-tune) | | | 61.7 | 203.39 | 7.56 | 30.0 | 199.83 | 6.73 |
| E2VID [44] (Fine-tune) | ✓ | | 61.1 | 203.40 | 7.45 | 30.9 | 199.44 | 6.80 |
| E2VID [44] (Fine-tune) | | ✓ | 61.5 | 202.23 | 7.56 | 29.4 | 203.97 | 6.08 |
| E2VID [44] (Scratch) | ✓ | | 21.3 | 362.08 | 2.43 | 4.9 | 333.13 | 1.36 |
| SSL-E2VID [37] (Pre-train) | | | 24.7 | 332.43 | 2.07 | 5.7 | 352.53 | 2.04 |
| SSL-E2VID [37] (Scratch) | | | 25.3 | 289.36 | 3.93 | 6.1 | 287.98 | 2.79 |
| Wang *et al.* [59] (Scratch) | | | 47.5 | 228.77 | 4.67 | 18.7 | 241.87 | 6.12 |
| Wang *et al.* [59] (Scratch) | ✓ | | 48.5 | 230.74 | 4.91 | 20.5 | 236.78 | 3.90 |
| Wang *et al.* [59] (Scratch) | | ✓ | 43.7 | 215.63 | 4.66 | 15.6 | 237.38 | 4.40 |
| Ours (Text Prompt) | | | 84.31 | 147.31 | 6.81 | 45.10 | 202.29 | 6.31 |
| Ours (Visual Prototype) | | | **85.56** | **107.58** | **9.35** | **51.50** | **174.24** | **7.32** |

Table 2: The results of event-based zero-shot object recognition using CLIP.

## 5.3. Zero-shot Object Recognition

We further demonstrate the advantages of our framework over supervised methods by enabling it to be used for zero-shot object recognition. The reconstructed images can be utilized for zero-shot object recognition thanks to CLIP's capacity to align images with any semantic concepts in an open vocabulary for zero-shot object recognition. To define "unseen" categories that are not seen during training, we use only 80 of the 100 categories for training and the remaining 20 for zero-shot testing.

In Table 2, we present the performance of zero-shot object recognition. Without any category labels, our methods with text prompt achieve a promising 84.31% and 45.10% on N-Caltech101 and N-ImageNet, respectively. This is 25.91% and 13.4% higher than E2VID pretrained with a large number of image pairs. These results validate that although our model learns through the process of being category-aware, generalization is possible for other categories, especially "unseen" in this case. Although FID and IS of our method using text prompts are relatively poor compared to pre-trained E2VID on N-ImageNet, this can also be overcome by using visual prototypes. Our method using the visual prototype is able to achieve a substantial improvement under all metrics. Specifically, in the N-Caltech101 dataset, our method increases the IS from 6.81 to 9.35, while in the N-ImageNet dataset, it increases the accuracy from 45.10 to 51.50.

## 6. Ablation Study and Dicussion

### 6.1. Ablation Study

We conduct an ablation study to understand how the components influence performance. Table 3 illustrates the effectiveness of each component on N-ImageNet dataset.

| | $\mathcal{L}_{att}$ | $\mathcal{L}_{rep}$ | $\mathcal{L}_{con}$ | RDS | | Accuracy ↑ |
|---|---|---|---|---|---|---|
| | | | | PPI | TRCI | |
| (1) | ✓ | | | | | 9.50 |
| (2) | ✓ | ✓ | | | | 14.00 |
| (3) | ✓ | ✓ | ✓ | | | 15.68 |
| (4) | ✓ | ✓ | | ✓ | | 28.42 |
| (5) | ✓ | ✓ | ✓ | ✓ | | 28.90 |
| (6) | ✓ | ✓ | ✓ | ✓ | ✓ | 28.91 |
| (7) | ✓ | ✓ | ✓ | ✓ | ✓ | 30.16 |

Table 3: Ablation study of the proposed method.

We start with a baseline, which denotes the removal of everything except attraction loss, $\mathcal{L}_{att}$. During the experiment, we notice that if we only apply attraction loss towards the network without repulsion loss, $\mathcal{L}_{rep}$, the collapse often takes place. Therefore, $\mathcal{L}_{rep}$ is essential for stable learning of $\mathcal{L}_{att}$. It is worth mentioning the impact of the proposed PPI on performance. Upon analyzing (4), which displays the results when PPI is added to (2), it can be seen that the accuracy has increased by 14.42. Upon examining (5) and (7), it can be observed that the accuracy increases by 1.26, validating the impact of the sub-sampling process via TRCI. Furthermore, the ablation results between (6) and (7) show that the reconstruction consistency between global and local preserves the spatial details of the image, leading to performance improvement. Finally, the model (7) that has all of its components incorporated achieves the highest performance.

### 6.2. Number of Samples in PPI

The $K$ used in PPI for selecting the sample with the highest probability impacts the performance of object recognition. Fig. 5 shows the evaluation of performance on N-ImageNet dataset based on different values of $K$. When $K$ is set to 32, which includes all samples in the batch, the
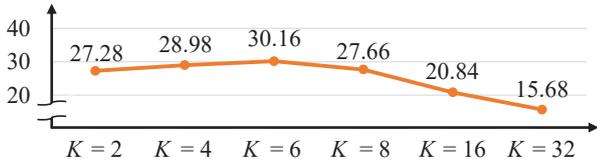
Figure 5: Object recognition accuracy on N-ImageNet (Mini) according to $K$ in PPI.
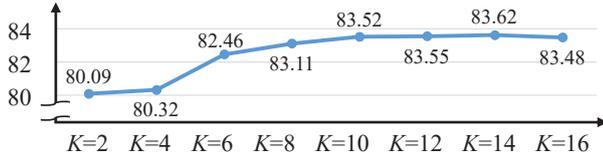


Figure 6: Object recognition accuracy on N-Caltech101 according to $K$ in PPI.

| | N-Caltech101 | | | N-ImageNet (Mini) | | |
|---|---|---|---|---|---|---|
| $\lambda_2 \setminus \lambda_3$ | 0.5 | 1 | 2 | 0.5 | 1 | 2 |
| 0.001 | 82.71 | 82.93 | 81.26 | 29.62 | 26.34 | 30.08 |
| 0.005 | 83.52 | 82.02 | 81.66 | 29.38 | 26.92 | 28.72 |
| 0.01 | 82.24 | 82.46 | 81.99 | 29.34 | 30.16 | 29.80 |
| 0.05 | 81.99 | 82.39 | 82.61 | 27.50 | 27.92 | 29.00 |
| 0.1 | 81.37 | 82.46 | 82.17 | 25.76 | 26.28 | 26.32 |

Table 4: The accuracy according to hyper-paramter in Eq. 11 when $\lambda_1 = 1$.

attraction loss, $\mathcal{L}_{att}$, is applied to even those samples with low reliability, resulting in the poorest performance. Due to the same reason, the performance is better when $K$ is 16 than when it is 32, but since unreliable samples are also selected, the accuracy is relatively lower than in other favorable situations. As $K$ decreases, the performance improves, but it starts to decline when $K$ is less than 6, indicating that even with high reliability, it is challenging to achieve optimal training if $\mathcal{L}_{att}$ is applied with very few samples. Similarly, we report the accuracy according to the $K$ value on the N-Caltech101 in Fig 6. Unless an extremely small number of samples are drawn, as the value of $K$ increases, the performance also improves. However, once a certain threshold is reached, there is little significant change, and it starts to decline with further increases. The accuracy exceeds 80% in all cases, showing robustness to $K$ values. Although the level of sensitivity varies depending on the dataset, we do confirm that satisfactory results can be obtained between $K = 4$ and $K = 10$.

## 6.3. Hyper-parameter Analysis

In Table 4, we report the accuracy according to $\lambda_2$ and $\lambda_3$ in Eq. 11. If the weight of the repulsion loss, $\lambda_2$, is set too high, the performance decreases. However, this can be mitigated by increasing the weight of the local-global re-
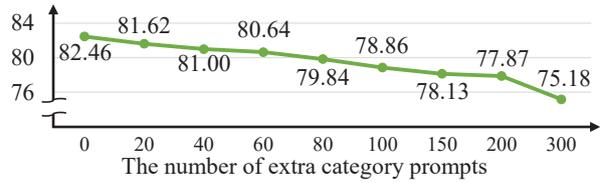


Figure 7: Object recognition accuracy on N-Caltech101 according to the number of extra category prompts.

construction consistency loss, $\lambda_3$. The results verify the robustness of our method according to the $\lambda$ on both datasets.

## 6.4. Learning with Superset Categories

In practical terms, it can be challenging for users to find corresponding categories of the event dataset. As a result, they often end up taking more categories as prompts. We tackle these issues by demonstrating that our joint learning framework works robustly even when employing more prompts than the categories of the event on the N-Caltech101 dataset. To avoid using additional prompts that are too irrelevant to the N-Caltech101 dataset, we randomly take additional prompts from N-ImageNet. We show the performance results according to the number of extra category prompts in Fig. 7. While the performance decreases with an increase in extra category prompts, it is worth noting that our algorithm remains reliable even when faced with over 100 additional prompts beyond the corresponding category. This suggests that our algorithm can function with superset categories, even in a situation when the number of extra categories is much larger than the actual categories.

## 7. Conclusion

In this work, we propose a joint learning framework for event-based object recognition with image reconstruction from events without labels and paired images. With the proposed losses and the reliable data sampling strategy, our approach allows stable learning of two tasks simultaneously from pseudo labels. As a result, our approach achieves exceptional performance in event-based object recognition and event-to-image reconstruction, all without requiring category labels and paired images. Our approach successfully addresses the challenge of label dependency in event-based research, which has been a significant obstacle. We anticipate that our work can be further extended to other tasks such as event-based zero-shot object detection, zero-shot semantic segmentation, etc., that have not explored so far in the research field for event cameras.

# References

[1] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7243–7252, 2017. 1, 2

[2] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 884–892, 2016. 2

[3] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based object classification for neuromorphic vision sensing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 491–501, 2019. 2

[4] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 136–152. Springer, 2020. 1, 2

[5] Hila Chefer, Sagie Benaim, Roni Paiss, and Lior Wolf. Image-based clip-guided essence transfer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 695–711. Springer, 2022. 2

[6] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. *arXiv preprint arXiv:2301.04926*, 2023. 2

[7] Wensheng Cheng, Hao Luo, Wen Yang, Lei Yu, Shoushun Chen, and Wei Li. Det: A high-resolution dvs dataset for lane extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[8] Hoonhee Cho, Jegyeong Cho, and Kuk-Jin Yoon. Learning adaptive dense event stereo from the image domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17797–17807, 2023. 2

[9] Hoonhee Cho, Jaeseok Jeong, and Kuk-Jin Yoon. Eomvs: Event-based omnidirectional multi-view stereo. *IEEE Robotics and Automation Letters*, 6(4):6709–6716, 2021. 2

[10] Hoonhee Cho and Kuk-Jin Yoon. Event-image fusion stereo using cross-modality feature propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 454–462, 2022. 2

[11] Hoonhee Cho and Kuk-Jin Yoon. Selection and cross similarity for event-image deep stereo. In *European Conference on Computer Vision*, pages 470–486. Springer, 2022. 2

[12] Gregory Cohen, Saeed Afshar, Garrick Orchard, Jonathan Tapson, Ryad Benosman, and Andre van Schaik. Spatial and temporal downsampling in event-based visual classification. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10):5030–5044, 2018. 1

[13] Matthew Cook, Luca Gugelmann, Florian Jug, Christoph Krautz, and Angelika Steger. Interacting maps for fast visual interpretation. In *The 2011 International Joint Conference on Neural Networks*, pages 770–776. IEEE, 2011. 2

[14] 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 conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[15] Li Fei-Fei, Rob Fergus, and Pietro Perona. 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*, pages 178–178. IEEE, 2004. 5

[16] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019. 3, 6

[17] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Asynchronous, photometric feature tracking using events and frames. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–765, 2018. 2

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[19] Yuhuang Hu, Hongjie Liu, Michael Pfeiffer, and Tobi Delbruck. Dvs benchmark datasets for object tracking, action recognition, and object recognition. *Frontiers in neuroscience*, 10:405, 2016. 2

[20] Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew J Davison. Simultaneous mosaicing and tracking with an event camera. *J. Solid State Circ*, 43:566–576, 2008. 2

[21] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 349–364. Springer, 2016. 2

[22] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2146–2156, 2021. 1, 2, 5, 6

[23] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with crossmodal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18032–18042, 2023. 2

[24] Simon Klenk, David Bonello, Lukas Koestler, and Daniel Cremers. Masked event modeling: Self-supervised pretraining for event cameras. *arXiv preprint arXiv:2212.10368*, 2022. 1, 2

[25] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2016. 1, 2, 6

[26] Jun Haeng Lee, Tobi Delbruck, and Michael Pfeiffer. Training deep spiking neural networks using backpropagation. *Frontiers in neuroscience*, 10:508, 2016. 1, 2

[27] Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11:309, 2017. 2

[28] Iulia-Alexandra Lungu, Federico Corradi, and Tobi Delbrück. Live demonstration: Convolutional neural network driven by dynamic vision sensor playing roshambo. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–1. IEEE, 2017. 2

[29] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5419–5427, 2018. 1, 6

[30] Nico Messikommer, Daniel Gehrig, Antonio Loquercio, and Davide Scaramuzza. Event-based asynchronous sparse convolutional networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 415–431. Springer, 2020. 1, 2

[31] Diederik Paul Moeys, Daniel Neil, Federico Corradi, Emmett Kerr, Philip Vance, Gautham Das, Sonya A Coleman, Thomas M McGinnity, Dermot Kerr, and Tobi Delbruck. Pred18: Dataset and further experiments with davis event camera in predator-prey robot chasing. *arXiv preprint arXiv:1807.03128*, 2018. 2

[32] Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. Phased lstm: Accelerating recurrent network training for long or event-based sequences. *Advances in neural information processing systems*, 29, 2016. 1, 2

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[34] Garrick Orchard, Ryad Benosman, Ralph Etienne-Cummings, and Nitish V Thakor. A spiking neural network architecture for visual motion estimation. In *2013 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 298–301. IEEE, 2013. 1, 2

[35] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015. 2, 5

[36] Garrick Orchard, Cedric Meyer, Ralph Etienne-Cummings, Christoph Posch, Nitish Thakor, and Ryad Benosman. Hfirst: A temporal approach to object recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):2028–2040, 2015. 1, 2

[37] Federico Paredes-Vallés and Guido CHE de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021. 2, 6, 8

[38] Paul KJ Park, Baek Hwan Cho, Jin Man Park, Kyoobin Lee, Ha Young Kim, Hyo Ah Kang, Hyun Goo Lee, Jooyeon Woo, Yohan Roh, Won Jo Lee, et al. Performance improvement of deep learning based gesture recognition using spa-

tiotemporal demosaicing technique. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1624–1628. IEEE, 2016. 1

[39] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2

[40] José Antonio Pérez-Carrasco, Bo Zhao, Carmen Serrano, Begona Acha, Teresa Serrano-Gotarredona, Shouchun Chen, and Bernabé Linares-Barranco. Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing–application to feedforward convnets. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2706–2719, 2013. 1, 2

[41] Stefano Pini, Guido Borghi, and Roberto Vezzani. Learn to see by events: Color frame synthesis from event and rgb cameras. *arXiv preprint arXiv:1812.02041*, 2018. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[43] Bharath Ramesh, Andrés Ussa, Luca Della Vedova, Hong Yang, and Garrick Orchard. Low-power dynamic object detection and classification with freely moving event cameras. *Frontiers in neuroscience*, 14:135, 2020. 2

[44] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3857–3866, 2019. 2, 6, 7, 8

[45] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 1, 2, 4, 5

[46] Christian Reinbacher, Gottfried Graber, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *arXiv preprint arXiv:1607.06283*, 2016. 2

[47] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 7

[48] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022. 2

[49] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V*, pages 308–324. Springer, 2019. 2

[50] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast im-

age reconstruction with an event camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 156–163, 2020. 2

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

[52] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow. In *European Conference on Computer Vision*, pages 628–645. Springer, 2022. 2

[53] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1731–1740, 2018. 1, 2

[54] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European Conference on Computer Vision*, pages 412–428. Springer, 2022. 2

[55] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928, 2022. 2

[56] Ajay Vasudevan, Pablo Negri, Bernabe Linares-Barranco, and Teresa Serrano-Gotarredona. Introduction and analysis of an event-based sign language dataset. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 675–682. IEEE, 2020. 2

[57] Lin Wang, Yujeong Chae, and Kuk-Jin Yoon. Dual transfer learning for event-based end-task prediction via pluggable event to image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2135–2145, 2021. 2

[58] Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 608–619, 2021. 2

[59] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10081–10090, 2019. 2, 6, 7, 8

[60] Lin Wang, Tae-Kyun Kim, and Kuk-Jin Yoon. Joint framework for single image reconstruction and super-resolution with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7657–7673, 2021. 1, 4, 5

[61] Yanxiang Wang, Bowen Du, Yiran Shen, Kai Wu, Guangrong Zhao, Jianguo Sun, and Hongkai Wen. Ev-gait: Event-based robust gait recognition using dynamic vision sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6358–6367, 2019. 6

[62] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2563–2572, 2021. 2

[63] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 2