This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

TESPEC: Temporally-Enhanced Self-Supervised Pretraining for Event Cameras

Mohammad Mohammadi [1, 2] Ziyi Wu [1,2] Igor Gilitschenski [1,2]
[1]University of Toronto [2]Vector Institute
{mohammadi, ziyiwu, gilitschenski}@cs.toronto.edu

Abstract

Long-term temporal information is crucial for event-based perception tasks, as raw events only encode pixel brightness changes. Recent works show that when trained from scratch, recurrent models achieve better results than feedforward models in these tasks. However, when leveraging self-supervised pre-trained weights, feedforward models can outperform their recurrent counterparts. Current self-supervised learning (SSL) methods for event-based pretraining largely mimic RGB image-based approaches. They pre-train feedforward models on raw events within a short time interval, ignoring the temporal information of events. In this work, we introduce TESPEC, a self-supervised pretraining framework tailored for learning spatio-temporal information. TESPEC is well-suited for recurrent models, as it is the first framework to leverage long event sequences during pre-training. TESPEC employs the masked image modeling paradigm with a new reconstruction target. We design a novel method to accumulate events into pseudo grayscale videos containing high-level semantic information about the underlying scene, which is robust to sensor noise and reduces motion blur. Reconstructing this target thus requires the model to reason about long-term history of events. Extensive experiments demonstrate our state-of-the-art results in downstream tasks, including object detection, semantic segmentation, and monocular depth estimation. Project webpage: https://mhdmohammadi.github.io/TESPEC_webpage.

1. Introduction

Event cameras are bio-inspired sensors that asynchronously record pixel intensity changes [26]. They offer distinct advantages, including low energy consumption, a high dynamic range, and high temporal resolution. There has been growing interest in applying event cameras to various computer vision tasks [4, 17, 30, 63, 95]. However, this novel data modality also poses unique challenges, e.g., the need for specialized models. Since individual events only encode short-term information, methods for complex tasks such as object detection usually aggregate events over a certain time interval [2, 9, 44]. Still, these *feedforward* models discard long-horizon history, making it hard to capture objects under small motion that trigger very few events. Recent models thus integrate recurrent modules [70] to utilize information beyond a relatively short time period [19, 29, 31, 50–52, 73, 98], demonstrating superior performance.

Despite recent progress, performance on many event vision tasks is limited by a lack of large labeled datasets [4, 30, 95]. In conventional RGB vision, self-supervised learning (SSL) has proven effective for improving performance in data-scarce scenarios [10, 35, 36]. Therefore, several studies have introduced SSL to event-based vision [48, 88, 89], aiming to pre-train robust feature extractors on large-scale, unlabeled event data. However, these approaches largely mimic conventional frame-based SSL. They aggregate events over short time intervals into 2D image-like representations, and then perform contrastive learning [89] or reconstruct masked locations [43, 48, 88]. A core limitation here is that real-world event data is notably sparser than RGB images. When converting events to frames, many pixels remain empty or contain merely noise, which fails to provide a meaningful learning signal [89]. Overall, pre-training models to extract long-term information from events is an underexplored research problem.

We address this gap with TESPEC, a self-supervised pretraining framework designed to learn long-term information from event sequences. Our approach follows the masked image modeling (MIM) paradigm [36, 86], where the model receives partially masked event streams and is trained to reconstruct an unmasked target. This requires the model to reason about current object locations based on their past motions. Unlike RGB images, event data has a distinctive temporal dimension. While raw events capture only low-level pixel brightness changes over short intervals, they can encode high-level semantic information when accumulated over longer periods. Prior studies have shown that high-quality grayscale videos can be reconstructed from raw event streams [67, 68]. Based on this insight, we propose using accumulated events to create pseudo-grayscale videos – resembling grayscale frames – as reconstruction targets. Compared to the short-term event frames used

in prior work, these pseudo videos provide richer long-term information and dense learning signals, supporting the training of recurrent architectures. In addition, approximating videos using events eliminates the need for paired video data in pre-training, making TESPEC a pure event-based self-supervised pre-training approach applicable to any event camera datasets.

A key challenge is how to obtain meaningful grayscale videos purely from event data. An intuitive solution is to use pre-trained event-to-frame reconstruction models [67, 68]. This is infeasible in our SSL setting as such models require paired event-video training data. An alternative approach would be using methods that estimate intensity signals by integrating events over time [5, 69]. However, we found their estimated videos suffer from severe motion blur when applied to outdoor sequences with high event rates. This is because these methods process each pixel separately and fail to adapt to the global motion of the scene. Object motion blur is particularly harmful to our pre-training approach, as it encourages the model to "remember" previous object positions, which contrasts the objectives of downstream perception tasks. To address this, we generalize the intensity estimation formulation from [69] to process events in global batches, enhancing robustness against local noise.

In summary, this work makes two main contributions: **(i)** We highlight an unexplored instance of event-based SSL that explicitly encourages the model to learn long-term temporal information, **(ii)** We propose an improved intensity video representation as the pre-training objective, leading to enhanced temporal information learning. As a consequence, our approach achieves state-of-the-art performance on multiple downstream event-based perception tasks.

2. Related Work

SSL for RGB images can be mostly categorized into three classes. Earlier works adopt the contrastive learning framework [1, 6, 10–14, 35, 37, 57, 74, 81]. The model is trained to perform instance discrimination by pushing two views of the same data closer, while separating views from different samples. Another line of work performs self-distillation [7, 32, 58, 93]. These methods only compare the model features on the same data, eliminating the need for negative samples. Recent years have witnessed the renaissance of reconstruction-based pre-training [38, 54, 60, 76, 91]. Inspired by the success of large language models [46, 64, 65], recent works propose masked image modeling (MIM) as the pre-training objective, where a model takes in a partially masked image, and is trained to reconstruct the unmasked one. Different reconstruction targets have been explored in the literature, including raw pixel values [8, 36, 86], discrete indices from a pre-trained tokenizer [3, 20, 62, 79], and intermediate features [15, 23, 80]. The representative work MAE [36] shows that ViTs [21] can reconstruct input images even with 75% of pixels masked out. In this work, we adopt the SSL framework similar to MAE and design a new reconstruction target tailored for event camera data.

MAE for RGB videos. Several works have extended the MAE framework to video data [24, 75, 78]. Compared to images, a distinct property of videos is temporal dynamics. Thus, some research incorporates motion into video MAE, such as guiding pixel masking with object motion [22, 41], predicting temporal differences between frames [71, 87], or predicting object trajectories [55, 72]. Interestingly, prior works observed degraded performance when only predicting frame differences [55, 71, 87]. Based on this insight, we propose reconstructing the accumulated event video instead of raw events as the pre-training objective. This forces the model to extract long-term temporal information from data.

SSL for event camera data. Due to a lack of large labeled datasets, many works have studied label-efficient learning on event data [16, 40, 56, 73, 82, 83, 92, 94]. Some design task-specific constraints for unsupervised learning [33, 42, 59, 77, 90, 96]. Instead, this paper aims to pre-train a versatile backbone. ECDP [89] trains the model to align features extracted from paired RGB images and events. MEM [48] and Huang et al. [43] follow MAE to reconstruct input events. The state-of-the-art ECDDP [88] designs sophisticated losses combining self-distillation and MIM. Notably, all these methods are designed for *feedforward* models, while *recurrent* models have shown better performances in the event vision literature [29, 50, 73, 98]. Our work is the first event SSL framework for pre-training recurrent models, which achieves state-of-the-art results on downstream tasks with a temporally-aware learning objective.

Recurrent architectures in event-based vision. To make a prediction, feedforward models only process events within a recent time range [2, 9, 44, 45]. However, raw event data encode pixel value changes and thus contain limited information. For tasks requiring long-term memory, recent works introduce recurrency to their backbones, including object detection [29, 49, 63], semantic segmentation [73], optical flow [31, 50], and depth estimation [19, 51, 52]. They usually add recurrent modules [39, 70] in the backbone, enabling reusing information from the last event segment. We follow this trend and propose TESPEC to learn a strong recurrent backbone to benefit these downstream tasks.

3. Method

TESPEC adopts the MAE [36] framework to pre-train a backbone for event camera data (Sec. 3.1). We construct intensity videos from raw event streams that contain rich temporal information (Sec. 3.2 & Sec. 3.3). The pre-training objective is a masked event reconstruction loss (Sec. 3.4). The overall pipeline of TESPEC is illustrated in Fig. 1.

3.1. Background

MAE and VideoMAE. MAE [36] performs masked image reconstruction with an asymmetric encoder-decoder archi-

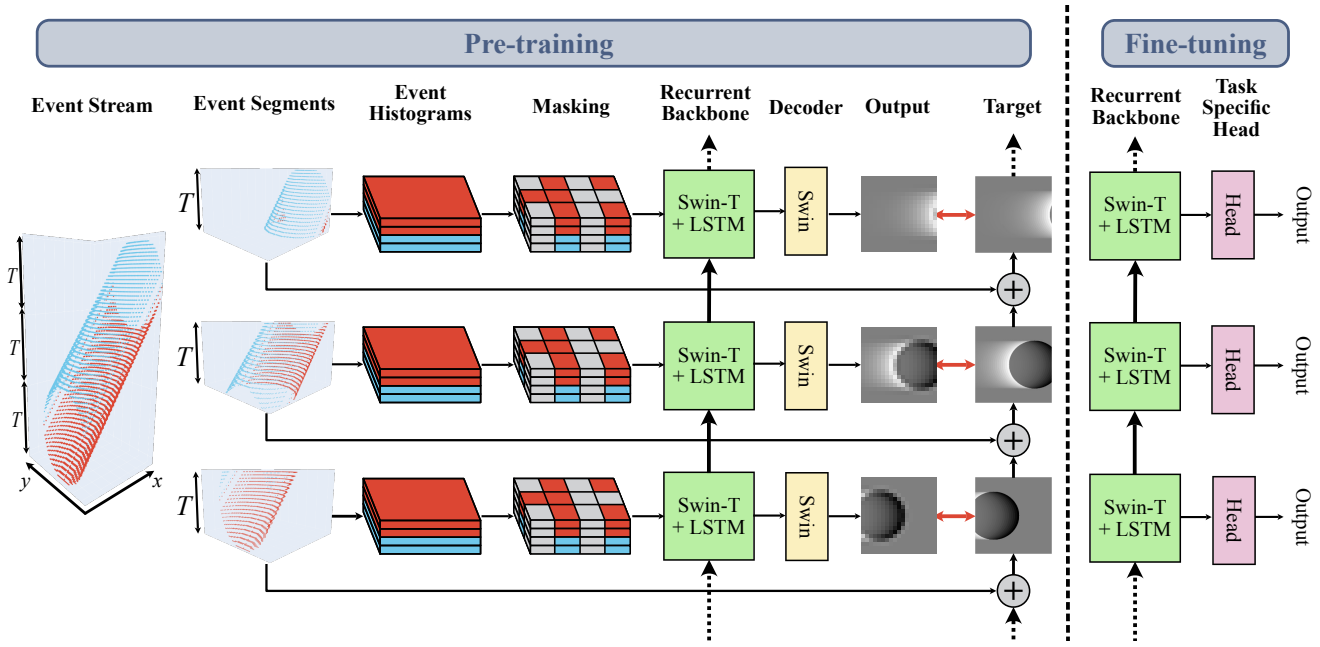


Figure 1. **Overview of the TESPEC Pipeline.** *Pre-training* (left): Given a raw event stream (red and blue represent negative and positive events, respectively), we first split it into non-overlapping segments, and convert each segment into an event histogram. They are used to update the estimated intensity video of the scene. Then, we apply temporal tube masking to event histograms that masks out the same spatial locations across time. The masked input is fed in our recurrent backbone, which extracts features by fusing history information from previous events and current inputs. With the extracted features, a lightweight feedforward decoder reconstructs the intensity video at masked patches. *Fine-tuning* (right): After pre-training, we discard the decoder and attach a new task-specific head to the recurrent backbone. Then, the whole model is fine-tuned on the downstream dataset.

tecture. Given an input image I, it is first divided into non-overlapping patches. Then, a high proportion of patches is randomly masked out, and the remaining ones are fed into a ViT [21] encoder to extract visual features. Finally, it appends equal amounts of learnable tokens to the encoded features and runs a ViT decoder to reconstruct masked patches. The training loss is a simple MSE over the masked patches:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\Omega|} \sum_{(x,y)\in\Omega} \left\| I(x,y) - \hat{I}(x,y) \right\|^2, \qquad (1)$$

where Ω is the subset of masked patches, and $\hat{I}$ is the reconstructed image. One key design choice in MAE is making the decoder significantly smaller than the encoder, leaving the encoder fully responsible for extracting visual features.

Later works extend MAE to video data [24, 75] with the same reconstruction-based learning objective. Compared to images, videos have a distinct property of temporal correlations between frames. Therefore, Tong et al. [75] propose tube masking, which masks out patches at the same spatial location throughout the entire video. This prevents information leaks and makes the pre-training task more challenging.

Event-Camera Data. Event cameras capture per-pixel log-intensity changes, and produce a stream of events $\mathcal{E} = \{e_i = (x_i, y_i, t_i, p_i)\}$. An event e_i is triggered at time t_i when the log-intensity at pixel (x_i, y_i) changes (since the last event at this location) beyond a pre-defined threshold C, i.e. when

$$L(x_i, y_i, t_i) - L(x_i, y_i, t_i - \Delta t) = p_i C, \qquad (2)$$

where $p_i \in \{-1, 1\}$ is the event polarity, and Δt is the time elapse since the last event at (x_i, y_i).

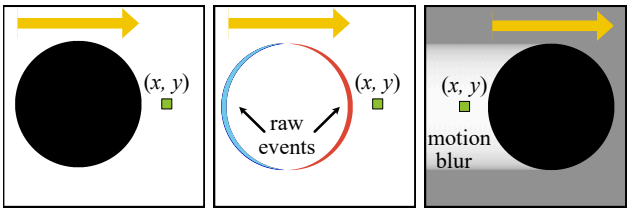
3.2. Event Processing for Temporal MAE

An MAE is not directly applicable to event data due to the event's sparse and asynchronous nature. To bridge the modality gap, we convert raw events to 2D frames. We adopt the event histogram representation [29] due to its simplicity and good performance in prior works [9, 44, 45]. We first split the event stream $\mathcal{E}$ into M non-overlapping segments $\{\mathcal{E}_i\}_{i=1}^{M}$, each covering events in a fixed time interval T. Then, we create a 4D tensor $\boldsymbol{s}_i \in \mathbb{R}^{2\times B\times H\times W}$ (dubbed stage) for each $\mathcal{E}_i$ as follows:

$$\boldsymbol{s}_i(p,\tau,x,y) = \sum_{e_j\in\mathcal{E}_i} \delta(p-p_j)\delta(x-x_j, y-y_j)\delta(\tau-\tau_j),$$
$$\tau_j = \left\lfloor \left(\frac{t_j}{T} - i + 1\right) \times B \right\rfloor, \qquad (3)$$

where δ is the Dirac delta function. Intuitively, each stage $\boldsymbol{s}_i$ further divides the event segment $\mathcal{E}_i$ into B temporal bins, and counts the number of positive and negative events separately in each pixel. Finally, we flatten the first two dimensions of $\boldsymbol{s}_i$, resulting in a multi-channel image-like representation with shape $(2B, H, W)$.

Vanilla event MAE. Given the 2D event frame $\boldsymbol{s}_i$, prior works [43, 48] simply mask it and run a feedforward model to do reconstruction. However, $\boldsymbol{s}_i$ only contains events in a short range and is usually sparse as shown in Fig. 3 (a).



(a) Toy example. (b) Triggered events. (c) Estimated intensity.

Figure 2. **Visualization of the estimated intensity image using Eq. (6) on a toy example.** (a) A black circle moves on a white background. (b) Assuming no sensor noise, it will trigger one negative and one positive event when passing through a pixel (x, y). (c) Due to the temporal decay term in Eq. (6), the events cannot cancel out, and thus will leave motion blur on the estimated image.

Reconstructing it is inefficient as many pixels are empty or contain only noisy events. In addition, events are mostly triggered at object boundaries. Forcing the model to reconstruct edges instead of the whole object may harm downstream dense prediction tasks such as depth estimation.

Temporally-enhanced event MAE. We aim to construct a representation from raw events that contain rich learning signals. Our key insight is that, while events in each time segment $\mathcal{E}_i$ present only low-level brightness changes, accumulated events over multiple segments can encode high-level semantic information. Indeed, while many objects could be static within a short segment (e.g., cars stop at the traffic light), they should move and trigger events when observed for longer. We thus propose integrating the entire event sequence $\mathcal{E}$ to obtain a video $\hat{L} \in \mathbb{R}^{H\times W\times MT}$ as:

$$\hat{L}(x, y, iT) = \mathrm{Int}(\hat{L}(x, y, (i-1)T), \mathcal{E}_i), \;\; \hat{L}(\cdot, \cdot, 0) = 0, \quad (4)$$

where $\mathrm{Int}(\cdot, \cdot)$ is the integration function to be discussed in Eq. (7). As we show in Sec. 3.3, $\hat{L}$ resembles the grayscale video of the scene and thus contains semantic information.

The multi-stage integration objective also aligns with the recurrent architectures in event vision tasks. These models typically take in event stages $\{\boldsymbol{s}_i\}_{i=1}^{M}$ sequentially, and use recurrent modules [70] to carry information over time:

$$F_i = \Phi_{\text{enc}}(\boldsymbol{c}_{i-1}, \boldsymbol{s}_i), \;\; \boldsymbol{c}_i = \Phi_{\text{rec}}(\boldsymbol{c}_{i-1}, \boldsymbol{s}_i), \;\; \boldsymbol{c}_0 = \mathbf{0}. \quad (5)$$

where Φ_{enc} is the backbone and F_i is the extracted feature, Φ_{rec} is the recurrent module and $\boldsymbol{c}_i$ is its memory state. Comparing Eq. (5) to Eq. (4), we find similar update rules between the output feature F_i and the accumulated video $\hat{L}_i$. Therefore, by reconstructing $\hat{L}_i$ from F_i, we can learn a backbone that extracts features with history information.

3.3. Improving the MAE Reconstruction Target

According to Eq. (2), given the initial grayscale image of a scene, one can reconstruct the real video L by integrating events per pixel over time. This is, however, impractical in our pure event-based SSL setting because it lacks image ground truth. Additionally, real-world event cameras often exhibit considerable sensor noise. Naive integration will quickly lead to visual quality degradation due to error accumulation [68]. Some studies thus incorporate temporal filtering to achieve robust intensity estimation from raw events [5, 69]. We build upon the formula in [69]:

$$\hat{L}(x_i, y_i, t_i) = \exp\left(-\alpha\Delta t\right) \cdot \hat{L}(x_i, y_i, t_i - \Delta t) + p_i C. \quad (6)$$

Intuitively, Eq. (6) updates the value of each pixel (x_i, y_i) individually. When a new event e_i comes at time t_i, it first performs an exponential decay on the previous state, and then increases or decreases the pixel value according to its polarity p_i. The decay term suppresses noise accumulation and is crucial to obtain a stable estimation.

However, the estimated image from Eq. (6) is prone to motion blur. Consider a toy example shown in Fig. 2 (a), where a black circle moves over a pixel (x, y) on a white background. When there is no sensor noise, the circle will only trigger one positive event and one negative event when it hits and leaves the pixel (Fig. 2 (b)). Without the decay term in Eq. (6), the positive event will first increase the pixel value by C, and then get canceled out by the negative event, resuming the original pixel value. Now with temporal decay, the two events cannot cancel out and the final pixel value becomes $(e^{-\alpha\Delta t} - 1) \cdot C$. As shown in Fig. 2 (c), this causes motion blur along the trajectory of the moving object and gets worse when the time elapse Δt is larger.

In real-world event streams, this is a common scenario, e.g., when a pedestrian walks across the street, the walking speed is small, and thus Δt is large, leading to severe motion blur as shown in Fig. 3 (b). Having motion blur in the reconstruction target is problematic, as it encourages the model to remember all past locations of the object. This is a waste of model capacity since downstream tasks often only focus on the current location of objects. In addition, real-world event cameras have sensor noise such as hot pixels that keep producing events of the same polarity, leading to severe noise on the estimated grayscale image.

Our solution. The issue with Eq. (6) is that each pixel is modeled separately. The intensity of a pixel is never updated if it does not receive new events, making the motion blur persist. To address it, we propose to update *all pixels* whenever the camera receives new events. Notably, in-



(a) Input events within a short time interval. (b) Estimated intensity, prior method (Eq. (6)). (c) Estimated intensity, our method (Eq. (7)).

Figure 3. **Visualization of the estimated intensity on a real-world example.** (a) Input events in a short interval are typically sparse. (b) The estimated image with the prior method contains motion blur behind the pedestrian, and severe sensor noise on the static background. (c) By considering events globally, our method produces images with sharp object motion and clean backgrounds.

stead of updating $\hat{L}$ on every incoming event, we assume all events within a temporal bin in $\mathcal{E}_i$ share the same timestamp, and update $\hat{L}$ based on each bin. This enables batch computation of our formula, which is fast on GPUs. Additionally, we adjust the exponential decay term with the number of received events. The new update rule is as follows:

$$\hat{L}(x,y,t) = \exp\left(-\alpha\Delta t \times \frac{n}{N}\right) \cdot \hat{L}(x,y,t-\Delta t) + E(x,y,t,\Delta t)\cdot C, \quad \text{where} \quad \Delta t = \frac{T}{B}. \tag{7}$$

Here, the time elapse Δt is always the interval of a temporal bin, $\frac{T}{B}$, $E(x,y,t,\Delta t)$ is the signed accumulation of events at pixel (x,y) within the temporal bin, and N is a fixed normalization factor. Importantly, the temporal decay factor is now positively correlated to the number of events n triggered during the entire Δt. Consider the example in Fig. 2 again, although the signed event count $E(x,y,t,\Delta t)$ remains zero after the circle leaves, the pixel value at (x,y) keeps getting decayed, thus eliminating the motion blur. This is verified by the real-world result in Fig. 3 (c), where objects and scene elements are much sharper. In addition, the noisy pixels on the background are also suppressed.

An interesting question is why involving the number of events n in the exponential decay term—even if we remove n from Eq. (7), the motion blur will still disappear over time. However, consider an extreme case where all objects are static and no events are triggered, $\hat{L}$ remains unchanged in Eq. (7) since $n = 0$, which is desired. In contrast, $\hat{L}$ will be decayed without considering n. Intuitively, the absence of new events suggests that the intensity should remain relatively unchanged, whereas a high number of events indicates sufficient information, allowing old events to be forgotten without compromising the estimation. Therefore, we use n to control the decay speed of history.

3.4. Event-based Temporal MAE Pre-training

Given an event stream $\mathcal{E}$, we first convert it to a sequence of stages $\{\mathcal{E}_i\}_{i=1}^{M}$ using Eq. (3) and divide each stage into patches. Then, we apply tube masking which masks out the same spatial patches across all stages in one training step, following VideoMAE [75]. The masked data is fed to the recurrent backbone sequentially to obtain features $\{F_i\}_{i=1}^{M}$ following Eq. (5). On the other hand, we construct a grayscale video $\hat{L}$ from $\mathcal{E}$ using Eq. (7). We leverage a lightweight feedforward decoder head Φ_{dec} that takes in event features extracted by the backbone, $\{F_i\}_{i=1}^{M}$, and predicts a video $\{\tilde{L}_i \in \mathbb{R}^{H\times W}\}_{i=1}^{M}$. The MSE between the estimated and the predicted video serves as our main loss,

$$\mathcal{L}_{\text{pretrain}} = \frac{1}{M\cdot|\Omega|}\sum_{i=1}^{M}\sum_{(x,y)\in\Omega}\left\|\tilde{L}_i(x,y) - \hat{L}(x,y,iT)\right\|^2, \quad \tilde{L}_i = \Phi_{\text{dec}}(F_i). \tag{8}$$

Method	Backbone	Recurrent	Pre-training	mAP ↑	
				Gen1	1Mpx
The best performance in the literature.					
FCCO [97]	SwinV2	No	MS-COCO	50.4	40.6
RVT [29]	MaxViT	Yes	-	47.2	47.4
GET-T [61]	Transformer	Yes	-	47.9	48.4
No Pre-training	Swin-T/7	No	-	36.2	34.2
RGB Supervised	Swin-T/7	No	ImageNet-1k	39.1	38.5
MoBY [85]	Swin-T/7	No	ImageNet-1k	41.0	36.9
No Pre-training	Swin-T/7	Yes	-	48.5	45.4
RGB Supervised	Swin-T/7	Yes	ImageNet-1k	50.5	48.9
MoBY [85]	Swin-T/7	Yes	ImageNet-1k	50.2	48.2
ECDDP [88]	Swin-T/7	No	E-TartanAir	40.8	38.5
ECDDP [88]	Swin-T/7	Yes	E-TartanAir	49.6	50.0
TESPEC (**Ours**)	Swin-T/7	Yes	1Mpx	**51.6**	**50.6**

Table 1. Object detection results on Gen1 [17] and 1Mpx [63] datasets. Best results are highlighted in bold. We report mean average precision (mAP) for evaluation. TESPEC achieves state-of-the-art results on both Gen1 and 1Mpx, surpassing all existing SSL methods in the RGB and event data domains.

A lightweight feedforward decoder forces the encoder to learn the spatio-temporal interactions in event sequences.

After pre-training, we discard the decoder, and attach a task-specific head to the recurrent backbone, which are jointly fine-tuned on the downstream dataset for each task.

4. Experiments

4.1. Experimental Setup

We evaluate TESPEC on three downstream tasks: object detection, semantic segmentation, and monocular depth estimation. Following prior works [88, 89], the pre-trained backbone is combined with a task-specific head and fine-tuned on the downstream datasets in a supervised learning manner. In this subsection, we only introduce the pre-training setting of our method, and leave the implementation details on each downstream task to its own subsection.

Pre-training dataset. We pre-train our backbone on the 1Mpx dataset [63] featuring outdoor driving scenarios. It contains around 15 hours of recordings with a 720×1280 resolution event camera [25], covering both day and night time. We chose it as 1Mpx has a higher resolution and more moving objects, thus more diverse motions compared to other event camera datasets [17, 30, 95].

Recurrent backbone. We adopt the Swin-Transformer architecture [53] with a patch size of 7 (dubbed Swin-T/7) as our encoder Φ_{enc} as it is the best-performing backbone in previous event-based SSL methods [88]. To make it a recurrent model, we add ConvLSTM [70] cells Φ_{rec} after each Swin-T stage. We take a much smaller version of Swin-T/7 containing only one block as our decoder Φ_{dec}. The decoder is a feedforward model without ConvLSTM. For the input event representation, we follow prior work [29] to aggregate events within every $T = 50$ ms into an event histogram, and each event histogram contains $B = 10$ tempo-

Method	Backbone	Recurrent	Pre-training	DSEC		DDD17	
				mIoU ↑	mAcc ↑	mIoU ↑	mAcc ↑
The best performance in the literature.							
ESS [73]	-	Yes	-	53.295	62.942	61.370	70.874
No Pre-training	Swin-T/7	No	-	53.665	61.194	53.996	64.591
RGB Supervised	Swin-T/7	No	ImageNet-1k	59.680	66.823	59.324	70.625
MoBY [85]	Swin-T/7	No	ImageNet-1k	58.553	66.070	57.667	67.639
No Pre-training	Swin-T/7	Yes	-	54.275	61.474	55.811	66.766
RGB Supervised	Swin-T/7	Yes	ImageNet-1k	60.867	68.258	60.800	70.670
MoBY [85]	Swin-T/7	Yes	ImageNet-1k	59.709	67.220	59.970	69.284
ECDP [89]	ResNet50	No	N-ImageNet	59.155	67.534	59.145	70.176
ECDDP† [88]	ResNet50	No	E-TartanAir	60.641	69.502	62.912	74.015
ECDDP† [88]	Swin-T/7	No	E-TartanAir	61.250	69.620	62.525	**74.301**
ECDDP‡ [88]	Swin-T/7	No	E-TartanAir	59.142	66.450	-	-
ECDDP [88]	Swin-T/7	Yes	E-TartanAir	59.820	67.680	59.748	71.056
TESPEC (**Ours**)	Swin-T/7	Yes	1Mpx	**62.774**	**70.612**	**65.187**	72.871

Table 2. Semantic segmentation results on DSEC [30] and DDD17 [4] datasets. Best results are highlighted in bold. We report mean interaction over union (mIoU) and mean class average (mAcc) for evaluation. TESPEC outperforms state-of-the-art methods in both metrics on DSEC and mIoU for DDD17, and achieves comparable results in mAcc on DDD17. † ECDDP leverages test-time augmentation to improve performance according to their official codebase. ‡ Our reproduced results by running the official codebase of ECDDP.

ral bins. Thanks to the Transformer architecture of Swin-T, our pre-trained backbone can be transferred to any downstream datasets with no resolution limit.

TESPEC pre-training details. One training sample contains a sequence of 15 event histograms. However, this only covers 0.75 seconds, which is still too short in many scenarios. Therefore, we sequentially sample the same event stream in consecutive training steps. At each step, we resume from the estimated grayscale video in the previous step and accumulate new input events. This enables us to train on minute-long event sequences. We use a masking ratio of 50%, the decay factor $\alpha = 5$, and a normalizing factor $N = 5{,}000$. Yet, as we will see in the ablation study (Sec. 4.5), our model is robust to these hyper-parameters. The entire model is pre-trained for 400k steps with a batch size of 8. The Adam optimizer [47] with a peak learning rate of 1×10^{-4} and 1.5k warmup steps is used.

Baselines. We compare our method with two groups of approaches: (i) state-of-the-art methods for each downstream task, and (ii) pre-trained weights from different domains:

- **RGB Supervised:** backbones that are supervised pre-trained on the ImageNet-1k classification dataset [18]. We take the official checkpoint from Swin-T/7 [53].
- **RGB SSL:** we take MoBY [85] which pre-trains Swin-T/7 on the ImageNet-1k dataset in a self-supervised way. We choose it as it is proposed by the authors of Swin-T.
- **Event SSL:** methods that perform unsupervised pre-training of the backbone on the event camera datasets, including ECDP [89] and the state-of-the-art ECDDP [88].

For tasks that are covered by previous works, we simply copy the numbers from their papers. Otherwise, we take their backbone and fine-tune them using our codebase under the same setting as ours to report results. The detailed setting of pre-training, and downstream task fine-tuning can be found in Appendix C.

4.2. Object Detection

Setting. We take the detection head design (YOLOX [27]) and the training codebase from a representative recurrent event-based object detector RVT [29]. We fine-tune all models on Gen1 [17] and 1Mpx [63] datasets with the default hyper-parameters in RVT.

Results. Tab. 1 presents the quantitative results. First, we observe that models with recurrent modules consistently outperforms their feedforward counterparts. This aligns with previous research that shows the importance of history information in the event-based detection task [29, 61]. Compared to no pre-training, pre-training on RGB data improves the performance slightly. Meanwhile, models pre-trained with TESPEC achieve substantial improvement of 3.1% and 5.2% in mAP on Gen1 and 1Mpx, respectively. This result surpasses ECDDP by 2.0% and 0.6% on both datasets. Moreover, our model consistently outperforms FCCO [97] on both datasets, despite FCCO's use of an optimized event representation and a model architecture 30% larger than ours (46M vs. 60M parameters). Overall, our results prove that TESPEC unleashes the power of recurrent models to learn long-term information from event sequences, which is crucial for the detection task.

4.3. Semantic Segmentation

Setting. Following prior work [88], we attach an UperNet head [84] to the backbone and fine-tune all models on the DSEC [30] and DDD17 [4] datasets. We train all models

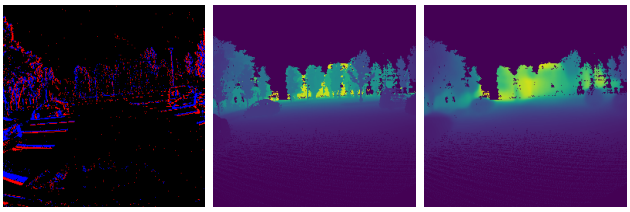
Method	Backbone	Recurrent	Pre-training	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	Abs $\downarrow$	RMS $\downarrow$	RMSlog $\downarrow$
The best performance in the literature.									
HMNet [34]	-	Yes	EventScape	0.588	0.784	0.889	4.171	7.534	0.397
PCDepth† [51]	-	Yes	EventScape	0.672	0.845	0.932	-	6.621	0.328
No Pre-training	Swin-T/7	No	-	0.345	0.731	0.853	5.585	8.671	0.476
RGB Supervised	Swin-T/7	No	ImageNet-1k	0.433	0.751	0.87	5.101	8.225	0.439
MoBY [85]	Swin-T/7	No	ImageNet-1k	0.5	0.753	0.868	4.974	8.066	0.433
No Pre-training	Swin-T/7	Yes	-	0.534	0.764	0.878	4.852	7.902	0.428
RGB Supervised	Swin-T/7	Yes	ImageNet-1k	0.596	0.799	0.906	4.116	7.174	0.375
MoBY [85]	Swin-T/7	Yes	ImageNet-1k	0.589	0.795	0.903	4.135	7.218	0.377
ECDP [89]	ResNet50	No	N-ImageNet	0.611	0.797	0.901	4.061	7.197	0.377
ECDDP [88]	ResNet50	No	E-TartanAir	0.612	0.809	0.915	3.889	6.805	0.359
ECDDP [88]	Swin-T/7	No	E-TartanAir	0.618	0.806	0.912	3.862	6.870	0.360
ECDDP [88]	Swin-T/7	Yes	E-TartanAir	0.563	0.794	0.903	4.191	7.291	0.383
TESPEC (**Ours**)	Swin-T/7	Yes	1Mpx	**0.634**	**0.830**	**0.926**	**3.690**	**6.654**	**0.343**

Table 3. Monocular depth estimation results on the MVSEC [95] dataset, with scores averaged across all testing sequences. Best results are highlighted in bold. We report threshold accuracy (δ_1, δ_2, and δ_3), absolute error (Abs), root mean squared error (RMS), and root mean squared logarithmic error (RMSlog) for evaluation. † PCDepth use both images and events as input. It only reports results on two sequences and does not release code. TESPEC outperforms all SSL and event-based methods, and achieve comparable results with PCDepth.



(a) Input events. (b) Ground-truth. (c) Prediction.

Figure 4. **Qualitative semantic segmentation results on DSEC.** We detect tiny objects such as traffic signs and street lights.



(a) Input events. (b) Ground-truth. (c) Prediction.

Figure 5. **Qualitative depth estimation results on MVSEC.** We predict accurate depth from sparse events thanks to recurrency.

for 50k steps with a learning rate of 2×10^{-4} and a linear warmup of 250 steps. Following ESS, the loss function is an unweighted sum of dice loss and cross-entropy loss. To match the label frequency, we set the event histogram duration T to 50ms for DSEC and 30ms for DDD17.

Results. Tab. 2 shows the quantitative results. Recurrent models always outperform their feedfoward counterparts. TESPEC surpasses baselines significantly in mIoU on both datasets. Specifically, we achieve 1.524% higher mIoU on DSEC and 2.275% higher mIoU on DDD17 compared to ECDDP. For mAcc, we are 0.992% higher on DSEC, and competitive on DDD17. Notably, ECDDP uses a segmentation head which is $3\times$ larger than ours. Overall, TESPEC brings a gain of 10.499% in mIoU and 9.138% in mAcc on DSEC, and 9.376% in mIoU and 6.105% in mAcc on DDD17 over no pre-training. This shows our method's effectiveness for pre-training recurrent models.

4.4. Monocular Depth Estimation

Setting. We attach the depth prediction head from MiDaS [66] to the backbone and fine-tune all models on the MVSEC [95] dataset. Following prior works [28, 88], we fine-tune on the "outdoor_day2" sequence and evaluate on the "outdoor_day1", "outdoor_night1", "outdoor_night2", and "outdoor_night3" sequences. All models are trained for 20k steps with a learning rate of 1×10^{-4}, a linear warm-up of 100 steps, and a batch size of 8. We use a weighted combination of a scale-invariant loss and a multi-scale scale-invariant gradient matching loss adopted from the state-of-the-art method HMNet [34].

Results. We summarize the quantitative results in Tab. 3. Due to the high sparsity of the MVSEC dataset, events within a single time segment is not enough for predicting accurate depth. Therefore, with randomly initialized weights or RGB pre-trained weights, recurrent models consistently outperform feedforward ones. Similarly, our method designed for recurrent architectures also performs better than ECDP and ECDDP across all metrics by a sizable margin. Compared to state-of-the-art event-based depth estimators in the literature that are supervised pre-trained on a synthetic dataset [28], TESPEC outperforms HMNet [34], and achieves comparable results with PCDepth [51] which takes in an additional image modality.

4.5. Ablation Study

We ablate each component of TESPEC and report results on the downstream DSEC semantic segmentation task.

Masking Ratio	DSEC	
	mIoU ↑	mAcc ↑
0%	55.619	62.982
25%	61.625	68.715
50%	**62.774**	**70.612**
75%	61.812	69.245

(a) **Masking Ratio.** TESPEC is robust to masking ratios between 25% and 75%, while autoencoding (0%) fails to learn useful representation.

Steps	DSEC	
	mIoU ↑	mAcc ↑
200k	61.554	68.924
300k	62.335	70.239
400k	**62.774**	**70.612**
500k	62.336	69.993

(b) **Number of Pre-training Steps.** The performance first grows with more training steps, and then saturates at 400k steps.

N	DSEC	
	mIoU ↑	mAcc ↑
500	61.742	69.072
5,000	**62.774**	**70.612**
50,000	61.428	68.857
200,000	60.059	67.658

(c) **Normalization Factor in Eq. (7).** A value of 5,000 balances the update of new events and the forgetting of old events.

Recon. Target	DSEC	
	mIoU ↑	mAcc ↑
Histograms	60.430	67.635
Eq. (6)	60.493	68.021
Eq. (7)	**62.774**	**70.612**

(d) **Reconstruction Target.** Our improved intensity video (Eq. (7)) outperforms both input reconstruction and prior work (Eq. (6)).

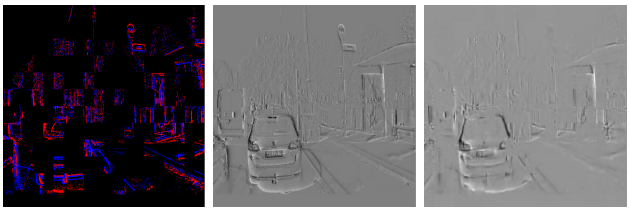
Dataset	DSEC	
	mIoU ↑	mAcc ↑
Gen1	60.128	67.149
1Mpx	**62.774**	**70.612**

(e) **Pre-training Dataset.** 1Mpx works better than Gen1 as it is of higher resolution, has more objects and diverse motion.

Normalization	DSEC	
	mIoU ↑	mAcc ↑
w/o norm	61.792	68.705
w/ norm	**62.774**	**70.612**

(f) **Patch Normalization.** Computing MSE loss with normalized patches improves the performance, which aligns with prior works [36].

Table 4. **Ablation study.** Our default settings are **bold**. We report downstream performance on DSEC semantic segmentation.



(a) Masked events. (b) Intensity video. (c) Reconstruction.

Figure 6. **Qualitative results on masked intensity video reconstruction.** Given sparse events that are only partially observable, TESPEC produces a holistic reconstruction of the scene.

Masking Ratio. The masking ratio determines the difficulty of the pre-training task. Tab. 4a shows that when not applying masking, i.e., doing autoencoding, the pre-training fails to learn useful representation. When the ratio is low, only a few patches are masked, providing limited signal for training. On the other hand, a high masking ratio discards too much information, making it challenging to learn. Given the sparse nature of event data, our optimal value 50% is lower than prior MAE work on RGB videos [75].

Number of Pre-training Steps. Tab. 4b shows that the performance first scales up with more pre-training steps, and then saturates at 400k steps.

Normalization Factor N in Eq. (7). This factor functions as a momentum value, balancing between retaining past information and updating with new inputs. A small N causes the model to forget prior events and only focus on recent inputs, hindering the learning of history information in recurrent models. A large N instead leads to high motion blur, slowing down the model's adaptation to new events. We chooses $N = 5{,}000$ which achieves the best balance.

Reconstruction Target. We compare three targets: (a) input event histograms, (b) estimated intensity videos from prior work [69] (Eq. (6)), and (c) our improved estimated intensity videos (Eq. (7)). Tab. 4d indicates that reconstructing event histograms leads to worse results. As discussed in Sec. 3.2, this is because input events only contain short-term information. In addition, using our intensity estimation as reconstruction target outperforms the naive one from prior work, highlighting the effectiveness of our modification. We show a reconstruction results from TESPEC in Fig. 6. Our model is capable of inpainting partially observable events thanks to the learned long-term history information.

Pre-training Dataset. Previous work [88] has shown that a pre-training dataset with diverse scenes and objects can enhance performance. We draw a similar observation in Tab. 4e. 1Mpx provides substantial gains compared to Gen1 due to its higher resolution and more diverse object motions.

Patch Normalization. Similar to MAE on RGB images, Tab. 4f shows that computing reconstruction loss in Eq. (8) on normalized ground-truth patches improves performance.

Pre-training Objective. To assess the effectiveness of MAE, we replace it with alternative contrastive loss functions and report the results in Appendix B.

5. Conclusion

In this paper, we present TESPEC, a pure event-based self-supervised pre-training framework. Our approach adopts the masked image modeling paradigm, and designs a reconstruction target that mines the long-term history information from raw event sequences. This design greatly enhances the performance of event-based recurrent models. Equipped with TESPEC, recurrent models can beat their feedforward counterparts on multiple event perception tasks.

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). We also acknowledge Vector Institute for computation support.

References

[1] Dosovitskiy Alexey, Philipp Fischer, Jost Tobias, Martin Riedmiller Springenberg, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE TPAMI*, 2016. 2

[2] Inigo Alonso and Ana C Murillo. Ev-SegNet: Semantic segmentation for event-based cameras. In *CVPR Workshops*, 2019. 1, 2

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: Bert pre-training of image transformers. In *ICLR*, 2022. 2

[4] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. DDD17: End-to-end davis driving dataset. *arXiv preprint arXiv:1711.01458*, 2017. 1, 6

[5] Christian Brandli, Lorenz Muller, and Tobi Delbruck. Real-time, high-speed video decompression using a frame-and event-based davis sensor. In *ISCAS*, 2014. 2, 4

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 2020. 2

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2

[8] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020. 2

[9] Nicholas FY Chen. Pseudo-labels for supervised learning on dynamic vision sensor data, applied to object detection under ego-motion. In *CVPRW*, 2018. 1, 2, 3

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 2

[11] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *NeurIPS*, 2020.

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

[13] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[14] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 2

[15] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *IJCV*, 2024. 2

[16] Hoonhee Cho, Hyeonseong Kim, Yujeong Chae, and Kuk-Jin Yoon. Label-free event-based object recognition via joint learning with image reconstruction from events. In *ICCV*, 2023. 2

[17] Pierre De Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*, 2020. 1, 5, 6

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[19] Anusha Devulapally, Md Fahim Faysal Khan, Siddharth Advani, and Vijaykrishnan Narayanan. Multi-modal fusion of event and rgb for monocular depth estimation using a unified transformer-based architecture. In *CVPRW*, 2024. 1, 2

[20] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, Nenghai Yu, and Baining Guo. PeCo: Perceptual codebook for bert pre-training of vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 552–560, 2023. 2

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2022. 2, 3

[22] David Fan, Jue Wang, Shuai Liao, Yi Zhu, Vimal Bhat, Hector Santos-Villalobos, Rohith MV, and Xinyu Li. Motion-guided masking for spatiotemporal representation learning. In *ICCV*, 2023. 2

[23] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 2

[24] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *NeurIPS*, 2022. 2, 3

[25] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Pooria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, et al. 5.10 a 1280× 720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86 μm pixels, 1.066 geps readout, programmable event-rate controller and compressive data-formatting pipeline. In *2020 IEEE International Solid-State Circuits Conference*, 2020. 5

[26] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *TPAMI*, 2020. 1

[27] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 6

[28] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 2021. 7

[29] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *CVPR*, 2023. 1, 2, 3, 5, 6

[30] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *RA-L*, 2021. 1, 5, 6

[31] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense optical flow from event cameras. In *3DV*, 2021. 1, 2

[32] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020. 2

[33] Jesse Hagenaars, Federico Paredes-Vallés, and Guido De Croon. Self-supervised learning of event-based optical flow with spiking neural networks. *NeurIPS*, 2021. 2

[34] Ryuhei Hamaguchi, Yasutaka Furukawa, Masaki Onishi, and Ken Sakurada. Hierarchical neural memory network for low latency event processing. In *CVPR*, 2023. 7

[35] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 2

[36] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 2, 8

[37] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *ICML*, 2020. 2

[38] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *NeurIPS*, 1993. 2

[39] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 2

[40] Yuhuang Hu, Tobi Delbruck, and Shih-Chii Liu. Learning to exploit multiple vision modalities by using grafted networks. In *ECCV*, 2020. 2

[41] Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao, and Limin Wang. Mgmae: Motion guided masking for video masked autoencoding. In *ICCV*, 2023. 2

[42] Ze Huang, Li Sun, Cheng Zhao, Song Li, and Songzhi Su. EventPoint: Self-supervised interest point detection and description for event-based camera. In *WACV*, 2023. 2

[43] Zhenpeng Huang, Chao Li, Hao Chen, Yongjian Deng, Yifeng Geng, and Limin Wang. Data-efficient event camera pre-training via disentangled masked modeling. *arXiv preprint arXiv:2403.00416*, 2024. 1, 2, 3

[44] Massimiliano Iacono, Stefan Weber, Arren Glover, and Chiara Bartolozzi. Towards event-driven object detection with off-the-shelf deep learning. In *IROS*, 2018. 1, 2, 3

[45] Zhuangyi Jiang, Pengfei Xia, Kai Huang, Walter Stechele, Guang Chen, Zhenshan Bing, and Alois Knoll. Mixed frame-/event-driven fast pedestrian detection. In *ICRA*, 2019. 2, 3

[46] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2

[47] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[48] Simon Klenk, David Bonello, Lukas Koestler, Nikita Araslanov, and Daniel Cremers. Masked event modeling: Self-supervised pretraining for event cameras. In *WACV*, 2024. 1, 2, 3

[49] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *TIP*, 2022. 2

[50] Yijin Li, Zhaoyang Huang, Shuo Chen, Xiaoyu Shi, Hongsheng Li, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. BlinkFlow: A dataset to push the limits of event-based optical flow estimation. In *IROS*, 2023. 1, 2

[51] Haotian Liu, Sanqing Qu, Fan Lu, Zongtao Bu, Florian Roehrbein, Alois Knoll, and Guang Chen. PCDepth: Pattern-based complementary learning for monocular depth estimation by best of both worlds. *arXiv preprint arXiv:2402.18925*, 2024. 2, 7

[52] Xu Liu, Jianing Li, Jinqiao Shi, Xiaopeng Fan, Yonghong Tian, and Debin Zhao. Event-based monocular depth estimation with recurrent transformers. *TCSVT*, 2024. 1, 2

[53] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5, 6

[54] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 2

[55] Yunyao Mao, Jiajun Deng, Wengang Zhou, Yao Fang, Wanli Ouyang, and Houqiang Li. Masked motion predictors are strong 3d action representation learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10191, 2023. 2

[56] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *RA-L*, 2022. 2

[57] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[58] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 2

[59] 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 *CVPR*, 2021. 2

[60] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2

[61] Yansong Peng, Yueyi Zhang, Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Get: Group event transformer for event-based vision. In *ICCV*, 2023. 5, 6

[62] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. BEiT v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2

[63] Etienne Perot, Pierre De Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *NeurIPS*, 2020. 1, 2, 5, 6

[64] Alec Radford. Improving language understanding by generative pre-training. 2018. 2

[65] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 2

[66] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 7

[67] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *CVPR*, 2019. 1, 2

[68] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *TPAMI*, 2019. 1, 2, 4

[69] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *ACCV*, 2018. 2, 4, 8

[70] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *NeurIPS*, 2015. 1, 2, 4, 5

[71] Yuxin Song, Min Yang, Wenhao Wu, Dongliang He, Fu Li, and Jingdong Wang. It takes two: Masked appearance-motion modeling for self-supervised video transformer pre-training. *arXiv preprint arXiv:2210.05234*, 2022. 2

[72] Xinyu Sun, Peihao Chen, Liangwei Chen, Changhao Li, Thomas H Li, Mingkui Tan, and Chuang Gan. Masked motion encoding for self-supervised video representation learning. In *CVPR*, 2023. 2

[73] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. ESS: Learning event-based semantic segmentation from still images. In *ECCV*, 2022. 1, 2, 6

[74] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 2

[75] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 2022. 2, 3, 5, 8

[76] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 2010. 2

[77] Jin Wang, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Unsupervised video deraining with an event camera. In *ICCV*, 2023. 2

[78] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *CVPR*, 2023. 2

[79] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[80] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 2

[81] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 2

[82] Ziyi Wu, Xudong Liu, and Igor Gilitschenski. EventCLIP: Adapting clip for event-based object recognition. *arXiv preprint arXiv:2306.06354*, 2023. 2

[83] Ziyi Wu, Mathias Gehrig, Qing Lyu, Xudong Liu, and Igor Gilitschenski. LEOD: Label-efficient object detection for event cameras. In *CVPR*, 2024. 2

[84] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 6

[85] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021. 5, 6, 7

[86] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *CVPR*, 2022. 1, 2

[87] Haosen Yang, Deng Huang, Bin Wen, Jiannan Wu, Hongxun Yao, Yi Jiang, Xiatian Zhu, and Zehuan Yuan. Self-supervised video representation learning with motion-aware masked autoencoders. *arXiv preprint arXiv:2210.04154*, 2022. 2

[88] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data dense pre-training. In *ECCV*, 2023. 1, 2, 5, 6, 7, 8

[89] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pre-training. In *ICCV*, 2023. 1, 2, 5, 6, 7

[90] Dehao Zhang, Qiankun Ding, Peiqi Duan, Chu Zhou, and Boxin Shi. Data association between event streams and intensity frames under diverse baselines. In *ECCV*, 2022. 2

[91] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2

[92] Xu Zheng and Lin Wang. EventDance: Unsupervised source-free cross-modal adaptation for event-based object recognition. In *CVPR*, 2024. 2

[93] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *ICLR*, 2022. 2

[94] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. EventBind: Learning a unified representation to bind them all for event-based open-world understanding. In *ECCV*, 2024. 2

[95] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *RA-L*, 2018. 1, 5, 7

[96] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*, 2019. 2

[97] Nikola Zubić, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. From chaos comes order: Ordering event rep-

resentations for object recognition and detection. In *ICCV*, 2023. 5, 6

[98] Nikola Zubic, Mathias Gehrig, and Davide Scaramuzza. State space models for event cameras. In *CVPR*, 2024. 1, 2