

# Event Camera Data Dense Pre-training

Yan Yang<sup>‡</sup>, Liyuan Pan<sup>†</sup>, and Liu Liu<sup>§</sup>

<sup>†</sup> School of CSAT, Beijing Institute of Technology, Beijing, China  
Corresponding author

<sup>‡</sup> BDSI, Australian National University, Canberra, Australia  
<sup>§</sup> KooMap Dept., Huawei, Beijing, China

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

**Abstract.** This paper introduces a self-supervised learning framework designed for pre-training neural networks tailored to dense prediction tasks using event camera data. Our approach utilizes solely event data for training.

Transferring achievements from dense RGB pre-training directly to event camera data yields subpar performance. This is attributed to the spatial sparsity inherent in an event image (converted from event data), where many pixels do not contain information. To mitigate this sparsity issue, we encode an event image into event patch features, automatically mine contextual similarity relationships among patches, group the patch features into distinctive contexts, and enforce context-to-context similarities to learn discriminative event features.

For training our framework, we curate a synthetic event camera dataset featuring diverse scene and motion patterns. Transfer learning performance on downstream dense prediction tasks illustrates the superiority of our method over state-of-the-art approaches.

**Keywords:** Event Camera Data · Self-supervised Learning · Dense Prediction

## 1 Introduction

An event camera asynchronously records pixel-wise brightness changes of a scene [18]. In contrast to conventional RGB cameras that capture all pixel intensities at a fixed frame rate, event cameras offer a high dynamic range and microsecond temporal resolution, and is robust to lighting changes and motion blur, showing promising applications in diverse vision tasks [4, 23, 43, 53].

This paper addresses the task of pre-training neural networks with event camera data for dense prediction tasks, including segmentation, depth estimation, and optical flow estimation. Our self-supervised method is pre-trained solely with event camera data. One can simply transfer our pre-trained model for dense prediction tasks. Please refer to Fig. 1 for the performance comparisons.

The direct way to pre-training is supervised training, using dense annotations for event data. However, due to the scarcity of dense annotations [4, 20, 53], training large-scale networks becomes challenging [12, 15].

An alternative to supervised pre-training is self-supervised learning for event camera data [47, 51], which has been proposed very recently. These approaches necessitate paired RGB images and event data, enforcing image-level embedding similarities between RGB images and event data. This form of RGB-guided pre-training directs networks to focus on the overall structure of events, neglecting intricate pixel-level features that are crucial for dense prediction tasks.

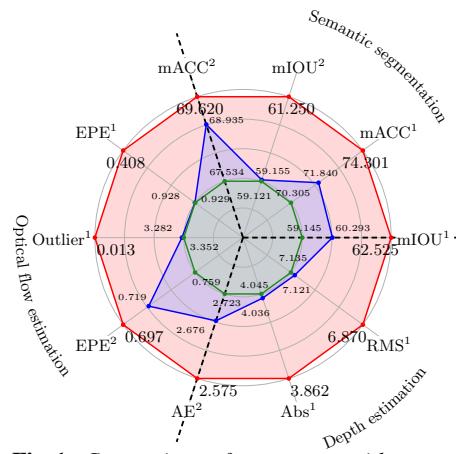
Next to pre-training is transferring the achievements of dense RGB pre-training [28, 42] to event camera data. One may first convert event camera data to an event image [47], split the image into patches, and then learn fine-grained patch features by enforcing patch-to-patch similarities in a self-supervised learning framework. While feasible, this baseline approach is constrained as event images are sparse, containing patches with little to no information, often from the meaningless background. The sparsity diminishes the discriminativeness of an event patch, introduces background noise/bias to the patch feature learning, and makes training unstable.

Inspired by the above discriminative self-supervised approaches that learn features at the image and patch level, we show that fine-grained event features can be learned by enforcing context-level similarities among patches. Our motivation is described below.

Given an event image, humans can recognize objects (*e.g.*, buildings and trees) by considering multiple similar pixels. In essence, a group of event image pixels contains sufficient information to make them discriminative. Inspired by this insight, we propose to automatically mine the contextual similarity relationship among patches, group patch features into discriminative contexts, and enforce context-to-context similarities. This context-level similarity, requiring no manual annotation, not only promotes stable training but also empowers the model to achieve highly accurate dense predictions.

Our contributions are summarized as follows:

1. A self-supervised framework for pre-training a backbone network for event camera dense prediction tasks. The pre-trained model can be transferred to diverse downstream dense prediction tasks;
2. Introduction of a context-level similarity loss to address the sparsity issue of event data for learning discriminative event features;
3. Construction of a pre-training dataset based on the TartanAir dataset [41], covering diverse scenes and motion patterns to facilitate network training;
4. State-of-the-art performance on standard event benchmark datasets for dense prediction tasks.



**Fig. 1:** Comparison of *our* scores with respect to the *second-best* and *third-best* scores for semantic segmentation [1, 4, 20, 39], optical flow estimation [20, 21, 53], and depth estimation [53]. Superscripts besides evaluation metrics are used to differentiate benchmark datasets for a specific task.

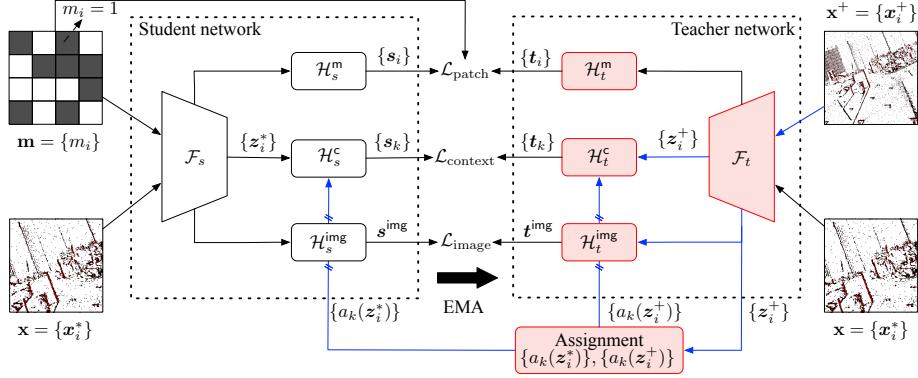
## 2 Related Works

We survey recent advancements in self-supervised learning frameworks applied to RGB and event image domains. We then provide an overview of event datasets used for network pre-training and downstream task fine-tuning.

**RGB image self-supervised learning.** Research in self-supervised learning generally falls into three categories: i) contrastive learning. Images are augmented into multiple views for instance discrimination. By defining a matching pair (*e.g.*, views from the same image), the similarity between them is maximized [9, 22]. Some works also enforce dissimilarity among non-matching pairs [7, 8, 10, 25, 45]; ii) masked image modeling. With unmasked image patches, the networks are trained to reconstruct masked ones. The reconstruction targets can be represented as intensity values of patch pixels [24, 46], discrete indices assigned by an image tokenizer [3, 16, 35], or patch embeddings obtained from pre-trained vision foundation models [17, 36]; iii) self-distillation. This category can be considered as an extension of contrastive learning from instances to groups [5, 6], and is usually combined with MIM [33, 52]. The similarity between matching image pairs is optimized by minimizing a cross-entropy loss, while MIM is optionally performed. For adapting self-supervised learning frameworks to dense prediction tasks, objectives at the patch/region level are proposed to maximize the similarity between matching patches [2, 28, 42, 48]. However, the spatial sparsity interferes with the patch-level objective and turns the network pre-training unstable, as most event image patches, containing little to no events, provide meaningless supervision signals.

**Event image self-supervised learning.** Explorations of self-supervised learning on event data remain in an early stage. Existing works [47, 51] primarily leverage a pre-trained CLIP network [36] and paired RGB images for training, guiding the event network to have similar outputs with the RGB network (*i.e.*, the image encoder of CLIP) in feature space. Because an event image is more similar to its paired RGB image at a high-level than at a low-level [49], these approaches concentrate on capturing the overall structures of the event image. This explains their substantial performance improvements in object recognition tasks for event data while lagging in various dense prediction tasks. In this paper, we do not require paired RGB images and pre-trained RGB networks, and focus on pre-training a versatile network by utilizing solely event data for diverse dense prediction tasks on event datasets.

**Event datasets.** Event cameras are bio-inspired sensors that pixel-wisely record spatial location, time, and polarity of brightness changes in a scene as an event sequence. One of the largest-scale event datasets covering diverse scenes is the N-ImageNet dataset [27]. It is built by moving an event camera to observe RGB images (from the ImageNet-1K dataset [14]) rendered by a monitor, and inherits scene diversity from the ImageNet-1K dataset. Existing event image self-supervised learning frameworks favor leveraging the N-ImageNet dataset for pre-training, enabling transfer learning for tasks such as object recognition [11, 27, 34, 38], depth estimation [53], semantic segmentation [4, 20], and optical flow estimations [20, 53]. This paper focuses on pre-training a network for the three dense prediction tasks. Moreover, considering the limited motion patterns in the N-ImageNet dataset [27], which are square, vertical, and horizontal, we curate a synthetic event dataset containing diverse motion patterns and scenes for pre-training.



**Fig. 2:** Overall architecture. During pre-training, our approach takes an event image  $\mathbf{x}^+$  and its affine-transformed counterpart  $\mathbf{x}^*$  as inputs, producing a pre-trained backbone network  $\mathcal{F}_s$ . A teacher network (colored by red boxes) and a student network are employed in the self-supervised training stage. Event images  $\mathbf{x}^+$  and  $\mathbf{x}^*$  are tiled into  $N$  patches, denoted as  $\mathbf{x}^+ = \{\mathbf{x}_i^+\}$  and  $\mathbf{x}^* = \{\mathbf{x}_i^*\}, i = 1, \dots, N$ . We randomly mask some patches of  $\mathbf{x}^*$  given to the student, but leave  $\mathbf{x}^*$  intact for the teacher. Patch-wise binary masks are represented by  $\mathbf{m} = \{m_i\}$ . Three similarity constraints are imposed based on output patch-wise features from the student and teacher backbones, respectively. They are: i) patch-level similarity. Patch-wise features of masked  $\mathbf{x}^*$  and  $\mathbf{x}^*$  are separately projected by heads  $\mathcal{H}_s^m$  in the student network and  $\mathcal{H}_t^m$  in the teacher network, obtaining embeddings  $\{s_i\}$  and  $\{t_i\}$ . To reconstruct masked patch embeddings, we employ a cross-entropy loss  $\mathcal{L}_{patch}$ ; ii) context-level similarity. Features  $\{z_i^+\}$  from the teacher network are assigned to  $K$  contexts, obtaining assignments  $\{a_k(z_i^+)\}$ .  $a_k(z_i^+)$  denotes the membership of the feature  $z_i^+$  to  $k$ -th context. The assignments of student features  $\{z_i^*\}$  are computed by directly transferring  $a_k(z_i^*)$  with an affine transformation. With the assignments  $\{a_k(z_i^+)\}$  and  $\{a_k(z_i^*)\}$ , we collect and pool all features assigned to each context using heads  $\mathcal{H}_s^c$  and  $\mathcal{H}_t^c$ , generating context embeddings  $\{s_k\}$  and  $\{t_k\}$ . A cross-entropy loss  $\mathcal{L}_{context}$  is used to learn masked context embeddings. The forward passes from  $\mathbf{x}^+$  are colored in blue, and the blocked lines mean crosslines; iii) image-level similarity.  $\{z_i^*\}$  and  $\{z_i^+\}$  are initially pooled separately and subsequently projected by the heads  $\mathcal{H}_s^{img}$  and  $\mathcal{H}_t^{img}$  into global image embeddings  $s^{img}$  and  $t^{img}$ . A cross-entropy loss  $\mathcal{L}_{image}$  is used to encourage image-level similarity.

### 3 Method

We present our self-supervised method in this section. Our network is trained end-to-end, and the overall architecture is shown in Fig. 2.

**Overall architecture.** We aim to learn discriminative features from event data for dense prediction tasks, such as optical flow estimation. Sharing similarities with the learning process of DINOv2 [33], we convert raw events to an image [54], and construct two event images  $\mathbf{x}^+$  and its augmentation  $\mathbf{x}^*$ . The two images are then fed into teacher and student networks to learn features, followed by enforcing similarities between the features of  $\mathbf{x}^+$  and  $\mathbf{x}^*$ . We enforce three types of feature similarities: i) patch-level similarity; ii) context-level similarity; iii) image-level similarity. Details of our components are provided below.

**Event image augmentations.** We perform a 2D affine transformation on  $\mathbf{x}^+$ , followed by GaussianBlur and ColorJitter [47], to create a distorted event image  $\mathbf{x}^*$ . We tile each image into  $N$  patches, i.e.,  $\mathbf{x}^+ = \{\mathbf{x}_i^+\}$  and  $\mathbf{x}^* = \{\mathbf{x}_i^*\}, i = 1, \dots, N$ . The linearity of the affine transformation establishes pixel correspondences between  $\mathbf{x}^+$  and  $\mathbf{x}^*$ . For each pixel in  $\mathbf{x}^*$ , we can find its corresponding pixel in  $\mathbf{x}^+$ , enabling context-level feature learning.

Image patches  $\{\mathbf{x}_i^+\}$  and  $\{\mathbf{x}_i^*\}$  are fed to the teacher and student networks for feature extraction. In the training stage, the student network is optimized by gradient descent. To avoid model collapse, the teacher network is kept as a momentum of the student network, and its parameters are updated with an exponential moving average (EMA) [25].

**Patch-level similarity.** We randomly mask some patches of  $\mathbf{x}^*$  given to the student, but leave  $\mathbf{x}^*$  intact for the teacher. The goal is to reconstruct masked patch embeddings, utilizing a cross-entropy loss between the patch features of both networks on each masked patch. This objective, introduced by [33], is briefly summarized below.

A patch-level binary mask  $\mathbf{m} = \{m_i\}, i = 1, \dots, N$  is randomly sampled. For  $\mathbf{x}_i^*$ , it is masked and replaced by a [MASK] token if  $m_i = 1$ . The unmasked patches and [MASK] tokens are fed to the student network  $\mathcal{F}_s$  to extract features, and a feature projection head  $\mathcal{H}_s^m$  is employed to obtain patch embeddings  $\{\mathbf{s}_i\} = \mathcal{H}_s^m(\mathcal{F}_s(\mathbf{x}^*, \mathbf{m}))$ .

Without masking, patches  $\{\mathbf{x}_i^*\}$  are fed to the teacher network  $\mathcal{F}_t$  to extract features, followed by a feature projection head  $\mathcal{H}_t$  to extract patch embeddings  $\{\mathbf{t}_i\} = \mathcal{H}_t(\mathcal{F}_t(\mathbf{x}^*))$ . The patch-level similarity objective is

$$\mathcal{L}_{\text{patch}} = \frac{1}{\|\mathbf{m}\|} \sum_{\substack{i=1 \\ m_i=1}}^N \text{CE}(\mathbf{t}_i, \mathbf{s}_i), \quad (1) \quad \text{CE}(\mathbf{t}, \mathbf{s}) = -\langle \mathcal{P}(\mathbf{t}), \log \mathcal{P}(\mathbf{s}) \rangle, \quad (2)$$

where  $\|\cdot\|$  is the L1 norm that computes the number of masked patches.  $\text{CE}(\cdot, \cdot)$  is the cross-entropy loss.  $\mathcal{P}(\cdot)$  is the Softmax function that normalizes the patch embedding to a distribution.  $\langle \cdot, \cdot \rangle$  is the dot product.

**Context-level similarity.** Reconstructing each masked patch embedding independently is prone to generating noisy embeddings. This is due to the sparsity of an event image. An event patch contains little information, and many patches are from a meaningless background (see Fig. 6). To overcome the limitations of independently reconstructing masked patch embeddings, we propose to mine contextual relationships among patch embeddings on the fly, and learn embeddings with context conditioning. We provide an overview in Fig. 3.

Specifically, we perform K-means clustering on patch features  $\{\mathbf{z}_i^+\} = \mathcal{F}_t(\mathbf{x}^+)$  of the teacher network, generating  $K$  cluster centers (i.e., contexts) and assignments  $a_k(\mathbf{z}_i^+)$ .  $a_k(\mathbf{z}_i^+)$  denotes the membership of the feature  $\mathbf{z}_i^+$  to  $k$ -th context, i.e., it is 1 if  $\mathbf{z}_i^+$  is closest to  $k$ -th context and 0 otherwise.

For each context, features assigned to it are aggregated by an attention pooling network [36], generating a context embedding  $\mathbf{s}_k$ . Collecting all context embeddings, we have embeddings  $\{\mathbf{s}_k\}, k = 1, \dots, K$ , describing features  $\mathcal{F}_t(\mathbf{x}^+)$  of the teacher.

For patch features  $\{z_i^*\} = \mathcal{F}_s(\mathbf{x}^*, \mathbf{m})$  of the student network, we use the same cluster centers. Due to the linearity of affine transformation, we can easily obtain the correspondence between patches  $\{x_i^*\}$  and  $\{x_i^+\}$ , and directly transfer the assignments  $\{a_k(z_i^+)\}$  to get  $\{a_k(z_i^*)\}$ . Given assignments  $a_k(z_i^*)$ , we follow the same pipeline to aggregate features  $\{z_i^*\}$  into context embeddings  $\{t_k\}$ ,  $k = 1, \dots, K$ .

By using adaptively mined contexts, such as roads and buildings, as proxies, we overcome the sparsity limitation of enforcing event patch-level similarity. In essence, we aim to enforce the similarity between a group of patches belonging to the same context. The context-level similarity loss  $\mathcal{L}_{\text{context}}$  is defined below

$$\mathcal{L}_{\text{context}} = \frac{1}{K} \sum_{k=1}^K \text{CE}(t_k, s_k) . \quad (3)$$

**Image-level similarity.** We aim to reconstruct masked image embedding of  $\mathbf{x}^*$ , by adding a cross-entropy loss between the image features of student and teacher networks on  $\mathbf{x}^*$  and  $\mathbf{x}^+$ .

Patch features  $\{z_i^*\}$  and  $\{z_i^+\}$  from the student and teacher network  $\mathcal{F}_s$  and  $\mathcal{F}_t$  are pooled and fed to feature projection heads  $\mathcal{H}_s^{\text{img}}$  and  $\mathcal{H}_t^{\text{img}}$ , generating image-level feature embeddings  $s^{\text{img}}$  and  $t^{\text{img}}$ , respectively. The image-level similarity objective is

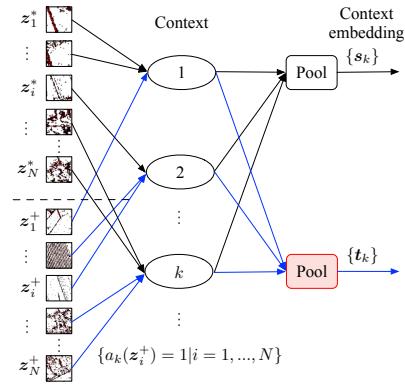
$$\mathcal{L}_{\text{image}} = \text{CE}(t^{\text{img}}, s^{\text{img}}) . \quad (4)$$

**Pre-training objective.** Our network is trained end-to-end. By using  $\lambda_1$  and  $\lambda_2$  hyper-parameters for balancing losses, we optimize the following objective,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{patch}} + \lambda_1 \mathcal{L}_{\text{context}} + \lambda_2 \mathcal{L}_{\text{image}} . \quad (5)$$

## 4 Experiments

**Pre-training dataset.** To pre-train our network, we synthesize an E-TartanAir event camera dataset from the TartanAir dataset [41]. The TartanAir dataset is collected in photo-realistic simulation environments, featuring various light conditions, weather, and moving objects. It has 1037 sequences with RGB frames of  $480 \times 640$  resolution. Different with N-ImageNet dataset [27] that has limited motion patterns, our E-TartanAir contains diverse motion patterns and scenes.



**Fig. 3: Context assignment and aggregation.** Given patch features  $\{z_i^*\}$  and  $\{z_i^+\}$ , we perform  $K$ -means clustering to mine  $K$  contexts, and obtain the patch-to-context assignments  $\{a_k(z_i^*)\}$  and  $\{a_k(z_i^+)\}$ , respectively. For the  $k$ -th context,  $\{z_i^+\}$  assigned to it  $\{a_k(z_i^+) = 1 | i = 1, \dots, N\}$  are pooled into a context embeddings  $t_k$ . Similarly,  $\{z_i^*\}$  are pooled into context embeddings  $s_k$ . The red box and blue lines denote components of our teacher network and forward passes of  $\{z_i^+\}$ , respectively.

**Table 1:** Comparison of semantic segmentation accuracies on the DDD17 [1, 4] and DSEC datasets [20, 39]. Mean interaction over union ( $mIoU$  (%)) and mean class accuracy ( $mA\overline{C}$  (%)) are used as evaluation metrics. ‘#Param’, ‘Pre. Dataset’, and ‘Pre. Epo.’ respectively denote the number of backbone parameters, pre-training dataset, and pre-training epoch.

Method	Backbone	#Param	Pre. Dataset	Pre. Epo.	DDD17		DSEC	
					$mIoU \uparrow$	$mA\overline{C} \uparrow$	$mIoU \uparrow$	$mA\overline{C} \uparrow$
<i>The best performance in the literature</i>								
ESS [39]	-	-	-	-	61.370	70.874	53.295	62.942
<i>Self-supervised ResNets.</i>								
SimCLR [7]	ResNet50	23M	ImageNet-1K	100	57.218	69.154	59.062	66.807
MoCo-v2 [8]	ResNet50	23M	ImageNet-1K	200	58.284	65.563	59.090	66.900
DenseCL [42]	ResNet50	23M	ImageNet-1K	200	57.969	71.840	59.121	68.935
ECDP [47]	ResNet50	23M	N-ImageNet	300	59.145	70.176	59.155	67.534
<b>Ours</b>	ResNet50	23M	E-TartanAir	300	<b>62.912</b>	<b>74.015</b>	60.641	69.502
<i>Self-supervised Transformers.</i>								
MoCo-v3 [10]	ViT-S/16	21M	ImageNet-1K	300	53.654	68.122	49.211	57.133
BeiT [3]	ViT-B/16	86M	ImageNet-1K	800	52.391	61.950	51.899	59.660
IBoT [52]	ViT-S/16	21M	ImageNet-1K	800	53.652	61.607	50.822	59.377
MAE [24]	ViT-B/16	86M	ImageNet-1K	800	53.758	64.783	51.958	59.839
SelfPatch [48]	ViT-S/16	21M	ImageNet-1K	300	54.287	62.821	51.475	59.164
DINOv2 [33]	ViT-S/16	21M	LVD-142M	-	53.846	64.500	52.165	59.795
CIM [29]	ViT-B/16	86M	ImageNet-1K	300	54.013	63.926	51.582	59.628
ECDP [47]	ViT-S/16	21M	N-ImageNet	300	54.663	66.077	52.517	60.553
ESViT [28]	Swin-T/7	28M	ImageNet-1K	300	60.293	70.305	56.517	63.798
Ours	ViT-S/16	21M	E-TartanAir	300	55.729	64.771	56.378	66.000
Ours	Swin-T/7	28M	E-TartanAir	300	62.525	74.301	<b>61.250</b>	<b>69.620</b>

**Implementation details.** We adopt ResNet50 [26], ViT-S/16 [15], and Swin-T/7 architectures as our backbones. The architectures of our projection heads follow [33, 36]. Our model is pre-trained for 300 epochs with batch size 1024. We set  $\lambda_1$  and  $\lambda_2$  to 0.1 and 0.9, respectively. The number of clusters is set to 8. Our code is available at [here](#).

**Baselines.** Our method is compared against two groups of methods: i) transfer learning of self-supervised pre-training. The initial weights of state-of-the-art methods are obtained in a self-supervised manner using the ImageNet-1K [14], N-ImageNet [27], or LVD-142M dataset [33]; ii) previous best. We compare with state-of-the-art methods specific to each downstream task, namely, semantic segmentation, flow estimation, and depth estimation. In tables, the symbols ‘ $\downarrow$ ’ and ‘ $\uparrow$ ’ indicate that a higher or lower value of a metric is preferable, respectively. The symbol ‘-’ denotes unavailability.

#### 4.1 Semantic Segmentation

**Settings.** Following the setup of [47], we evaluate on the DDD17 [1, 4] and DSEC dataset [20, 39] for semantic segmentation. The two datasets contain selected intervals of multiple event sequences, covering 6 and 11 semantic classes, respectively.

**Table 2:** Comparison of optical flow estimation accuracies on the MVSEC dataset [53]. End-point error (EPE) and outlier ratios (%) [47] are used as evaluation metrics. Pixels with EPE above 3 and 5% of the ground truth optical flow magnitudes are deemed as outliers [32].

Method	Backbone	<i>indoor_flying1</i>		<i>indoor_flying2</i>		<i>indoor_flying3</i>	
		EPE↓	Outlier↓	EPE↓	Outlier↓	EPE↓	Outlier↓
<i>The best performance in the literature</i>							
DCEIFlow [40]	-	0.748	0.597	1.388	8.015	1.132	5.294
<i>Self-supervised ResNets.</i>							
SimCLR [7]	ResNet50	0.646	0.488	1.445	9.331	1.188	5.507
MoCo-v2 [8]	ResNet50	0.612	0.459	1.359	8.683	1.130	5.201
ECDP [47]	ResNet50	0.604	0.354	1.352	8.572	1.122	5.263
DenseCL	ResNet50	0.634	0.529	1.349	7.596	1.130	5.176
Ours	ResNet50	0.413	0.055	0.489	0.041	0.462	0.007
<i>Self-supervised Transformers.</i>							
MoCo-v3 [10]	ViT-S/16	0.648	0.744	1.361	8.660	1.119	5.594
BeiT [3]	ViT-B/16	0.613	0.438	1.159	5.622	1.013	4.654
iBoT [52]	ViT-S/16	0.630	0.562	1.259	6.752	1.038	4.912
MAE [24]	ViT-B/16	0.613	0.167	1.293	6.952	1.109	4.635
SelfPatch [48]	ViT-S/16	0.623	0.317	1.337	7.894	1.097	5.286
DINOv2 [33]	ViT-S/16	0.602	0.325	1.196	6.185	0.990	4.333
CIM [29]	ViT-B/16	0.625	0.491	1.332	8.926	1.040	4.869
ECDP [47]	ViT-S/16	0.614	0.046	1.261	6.689	1.001	3.111
ESViT [28]	Swin-T/7	0.812	1.224	1.338	8.316	1.078	5.185
Ours	ViT-S/16	0.508	0.112	0.691	0.290	0.610	0.075
Ours	Swin-T/7	<b>0.362</b>	<b>0.035</b>	<b>0.445</b>	<b>0.002</b>	<b>0.417</b>	<b>0.001</b>

**Table 3:** Comparisons of optical flow estimation accuracies on the DSEC dataset [20]. Note that IDNet, ranking first previously, maintains anonymity at the time of submission. According to the DSEC leaderboard, we present results for 1/2/3-pixel error (1/2/3-PE), end-point error (EPE), and angular error (AE). All data is sourced from the online benchmark at the time of submission.

Methods	1PE↓	2PE↓	3PE↓	EPE↓	AE↓
E-RAFT [21]	12.742	4.740	2.684	0.788	2.851
MultiCM [37]	76.570	48.480	30.855	3.472	13.983
E-Flowformer [30]	11.225	4.102	2.446	0.759	2.676
TMA [31]	10.863	3.972	2.301	0.743	2.684
OF_EV_SNN [13]	53.671	20.238	10.308	1.707	6.338
IDNet [44]	10.069	3.497	2.036	0.719	2.723
Ours (ResNet50)	9.013	3.290	1.983	0.701	2.611
Ours (ViT-S/16)	9.288	3.339	2.005	0.714	2.615
Ours (Swin-T/7)	<b>8.887</b>	<b>3.199</b>	<b>1.958</b>	<b>0.697</b>	<b>2.575</b>

**Results.** Tab. 1 gives the comparisons on the DDD17 and DSEC datasets. Consistently, our method surpasses the state-of-the-art methods within the backbone groups of ResNet50, ViT-S/16, and Swin-T/7, and achieves a better performance than the methods

**Table 4:** Comparison of depth estimation accuracies on the MVSEC dataset [53]. Averaged scores across all sequences with a cutoff threshold at 30 meters are reported. Threshold accuracy ( $\delta_1$ ,  $\delta_2$ , and  $\delta_3$ ), absolute error (Abs), root mean squared error (RMS), and root mean squared logarithmic error (RMSlog) are used as evaluation metrics. The inputs of HMNet<sup>1</sup> are events, and HMNet<sup>2</sup> additionally takes RGB frames as inputs.

Method	Backbone	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	Abs $\downarrow$	RMS $\downarrow$	RMSlog $\downarrow$
<i>The best performance in the literature.</i>							
HMNet <sup>1</sup> [23]	-	0.626	0.818	0.912	2.882	4.772	0.361
HMNet <sup>2</sup> [23]	-	0.628	0.803	0.905	2.908	4.858	0.359
<i>Self-supervised ResNets.</i>							
SimCLR [7]	ResNet50	0.633	0.822	0.918	2.886	4.612	0.351
MoCo-v2 [8]	ResNet50	0.647	0.827	0.919	2.817	4.556	0.346
ECDP [47]	ResNet50	0.651	0.829	0.921	2.798	4.530	0.343
DenseCL [42]	ResNet50	0.649	0.826	0.920	2.813	4.541	0.344
Ours	ResNet50	0.649	<b>0.837</b>	<b>0.931</b>	2.713	4.302	<b>0.330</b>
<i>Self-supervised Transformers.</i>							
MoCo-v3 [10]	ViT-S/16	0.630	0.814	0.909	3.043	4.817	0.362
BeiT [3]	ViT-B/16	0.622	0.805	0.903	3.147	4.965	0.372
iBoT [52]	ViT-S/16	0.623	0.816	0.912	2.998	4.736	0.360
MAE [24]	ViT-B/16	0.612	0.802	0.900	3.214	5.075	0.377
SelfPatch [48]	ViT-S/16	0.605	0.801	0.900	3.435	5.067	0.380
DINOv2 [33]	ViT-S/16	0.612	0.805	0.903	3.181	5.030	0.375
CIM [29]	ViT-B/16	0.625	0.808	0.904	3.108	4.906	0.370
ECDP [47]	ViT-S/16	0.614	0.802	0.899	3.228	5.104	0.378
ESViT [28]	Swin-T/7	0.644	0.829	0.923	2.796	4.482	0.342
Ours	ViT-S/16	0.649	0.827	0.920	2.815	4.476	0.343
Ours	Swin-T/7	<b>0.658</b>	<b>0.837</b>	0.928	<b>2.658</b>	<b>4.257</b>	<b>0.330</b>

pre-trained with a larger ViT-B/16 backbone. For example, our method with a Swin-T/7 backbone achieves mIoU/mACC scores at 62.525%/74.301% and 61.250%/69.620% on the DDD17 and DSEC datasets, respectively, outperforming all other methods. Even though DINOv2 [33] is trained on the huge LVD-142M dataset, our method significantly outperforms it.

## 4.2 Flow Estimation

**Settings.** We compare our method with state-of-the-art methods on the MVSEC dataset [53]. End-point error (EPE) and outlier ratios (%) are used as evaluation metrics [40, 47]. In accordance with [47], the evaluations are performed on the ‘indoor\_flying1’, ‘indoor\_flying2’, and ‘indoor\_flying3’ sequences.

Additionally, our method is evaluated on the DSEC-Flow benchmark<sup>4</sup> [20, 21], securing the first-place position at the time of submission.

<sup>4</sup> <https://dsec.ifi.uzh.ch/uzh/dsec-flow-optical-flow-benchmark/>

**Table 5:** Comparison of depth estimation accuracies on the MVSEC dataset [53]. Averaged scores across all sequences are reported. Threshold accuracy ( $\delta_1$ ,  $\delta_2$ , and  $\delta_3$ ), absolute error (Abs), root mean squared error (RMS), and root mean squared logarithmic error (RMSlog) are used as evaluation metrics. The inputs of HMNet<sup>1</sup> are events, and HMNet<sup>2</sup> additionally takes RGB frames as inputs.

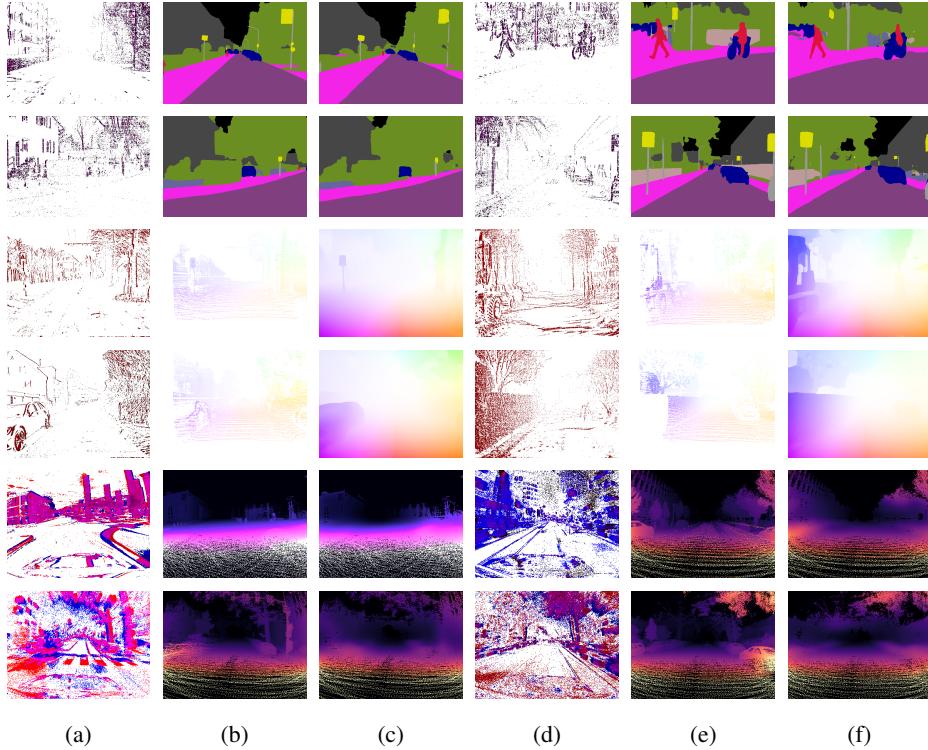
Method	Backbone	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	Abs $\downarrow$	RMS $\downarrow$	RMSlog $\downarrow$
<i>The best performance in the literature.</i>							
HMNet <sup>1</sup> [23]	-	0.588	0.784	0.889	4.171	7.534	0.397
HMNet <sup>2</sup> [23]	-	0.582	0.754	0.860	4.614	8.602	0.430
<i>Self-supervised ResNets.</i>							
SimCLR [7]	ResNet50	0.594	0.789	0.897	4.176	7.343	0.386
MoCo-v2 [8]	ResNet50	0.609	0.797	0.901	4.045	7.135	0.377
ECDP [47]	ResNet50	0.611	0.797	0.901	4.061	7.197	0.377
DenseCL [42]	ResNet50	0.610	0.798	0.903	4.036	7.121	0.375
Ours	ResNet50	0.612	<b>0.809</b>	<b>0.915</b>	3.889	<b>6.805</b>	<b>0.359</b>
<i>Self-supervised Transformers.</i>							
MoCo-v3 [10]	ViT-S/16	0.590	0.782	0.891	4.313	7.466	0.394
BeiT [3]	ViT-B/16	0.584	0.775	0.886	4.398	7.562	0.402
iBoT [52]	ViT-S/16	0.583	0.782	0.892	4.309	7.521	0.394
MAE [24]	ViT-B/16	0.575	0.772	0.884	4.449	7.601	0.405
SelfPatch [48]	ViT-S/16	0.567	0.768	0.882	4.515	7.735	0.410
DINOv2 [33]	ViT-S/16	0.575	0.774	0.885	4.449	7.653	0.406
CIM [29]	ViT-B/16	0.585	0.777	0.888	4.356	7.495	0.398
ECDP [47]	ViT-S/16	0.576	0.772	0.883	4.491	7.680	0.406
ESViT [28]	Swin-T/7	0.604	0.796	0.903	4.083	7.219	0.377
Ours	ViT-S/16	0.610	0.800	0.906	3.987	<b>6.957</b>	0.369
Ours	Swin-T/7	<b>0.618</b>	0.806	0.912	<b>3.862</b>	6.870	0.360

**Results.** Tab. 2 presents the comparisons on MVSEC dataset. Among the three different backbone groups, we have the most accurate flow estimation by pre-training with a Swin-T/7 backbone, and the EPE and outlier ratios on the three sequences are 0.362/0.035%, 0.445/0.002%, and 0.417/0.001%, respectively, which are significantly better than all other methods.

Results for the DSEC-Flow benchmark are given in Tab. 3. Compared with the method IDNet [44], previously holding the top position, our method achieves superior optical flow estimation accuracy. For example, our method with a ResNet50, ViT-S/16, and Swin-T/7 backbone respectively improves the state-of-the-art EPE/AE scores from 0.719/2.723 to 0.701/2.611, 0.714/2.615, and 0.697/2.575.

### 4.3 Depth Estimation

**Settings.** We evaluate the performance of our methods for depth estimation on the MVSEC dataset [53]. Following [19], the evaluations are performed on the ‘outdoor\_day1’, ‘outdoor\_night1’, ‘outdoor\_night2’, and ‘outdoor\_night3’ sequences.



**Fig. 4:** Qualitative comparison examples of dense predictions, namely, semantic segmentation (1<sup>st</sup>-2<sup>nd</sup> rows), optical flow estimation (3<sup>rd</sup>-4<sup>th</sup> rows), and depth estimation (5<sup>th</sup>-6<sup>th</sup> rows). (a) and (d): event images. Red and blue pixels depict positive and negative events, respectively. (b) and (e): ground-truth labels. (c) and (f): our model predictions. The brightness of depth maps in the 5<sup>th</sup> row of (b) and (c) is enhanced for visualization.

**Results.** The comparisons of our methods and state-of-the-art methods with and without a cutoff threshold at 30 meters are given in Tab. 4 and Tab. 5, respectively. Though the previous best method HMNet [23] performs supervised pre-training using ground-truth depth before fine-tuning on the MVSEC dataset, all our methods outperform it. For example, in Tab. 5, the averaged root mean squared error of HMNet is 7.534, while the errors of our methods with ResNet50, ViT-S/16, and Swin-T/7 backbones are 6.805, 6.957, and 6.870, respectively.

Sample prediction results of our method on the semantic segmentation, optical flow estimation, and depth estimation tasks are provided in Fig. 4.

#### 4.4 Discussions

We perform ablations on the DSEC semantic segmentation dataset [20, 39] to study our model components. We set the pre-training backbone and dataset to the Swin-T/7 and E-TartanAir dataset, except where otherwise indicated.

**Table 6:** (a)-(c) Comparison of state-of-the-art methods pre-trained on the N-imageNet datasets with backbones of ResNet50, ViT-S/16, Swin-T/7. (d)-(f) Comparison of state-of-the-art methods pre-trained on the E-TartanAir datasets with backbones of ResNet50, ViT-S/16, Swin-T/7.

Method	mIOU↑	mACC↑	Method	mIOU↑	mACC↑	Method	mIOU↑	mACC↑
SelfPatch	57.881	64.916	SelfPatch	50.442	58.452	SelfPatch	52.997	59.928
ESViT	57.796	64.928	ESViT	51.011	58.902	ESViT	53.051	60.094
ECDP	59.155	67.534	ECDP	52.517	60.553	ECDP	55.842	63.548
Ours	<b>60.243</b>	<b>69.195</b>	Ours	<b>54.897</b>	<b>62.527</b>	Ours	<b>56.654</b>	<b>65.250</b>

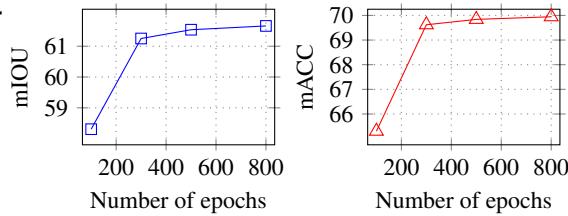
(a) All methods are pre-trained using the ResNet50 backbone, on the N-ImageNet dataset.	(b) All methods are pre-trained using the ViT-S/16 backbone, on the N-ImageNet dataset.	(c) All methods are pre-trained using the Swin-T/7 backbone, on the N-ImageNet dataset.						
(d) All methods are pre-trained using the ResNet50 backbone, on the E-TartanAir dataset.	(e) All methods are pre-trained using the ViT-S/16 backbone, on the E-TartanAir dataset.	(f) All methods are pre-trained using the Swin-T/7 backbone, on the E-TartanAir dataset.						
Method	mIOU↑	mACC↑	Method	mIOU↑	mACC↑	Method	mIOU↑	mACC↑
SelfPatch	58.365	65.180	SelfPatch	52.347	59.947	SelfPatch	57.243	66.070
ESViT	59.058	65.879	ESViT	51.945	60.470	ESViT	58.593	65.746
ECDP	59.572	68.317	ECDP	53.229	61.712	ECDP	56.568	64.234
Ours	<b>60.641</b>	<b>69.502</b>	Ours	<b>55.729</b>	<b>64.771</b>	Ours	<b>61.250</b>	<b>69.620</b>

### Pre-training datasets and backbones.

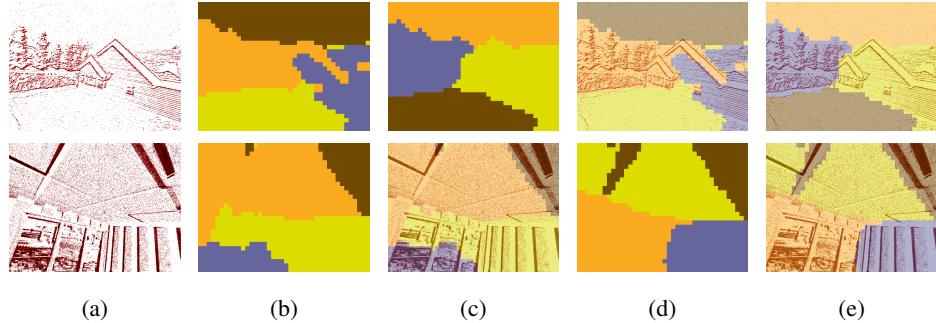
We investigate the impact of different pre-training datasets and backbones. We pre-train state-of-the-art methods with backbones of ResNet50, ViT-S/16, and Swin-T/7 respectively on the N-ImageNet and E-TartanAir datasets. The results are given in Tab. 6. We

have the following observations: i) Our proposed E-TartanAir dataset outperforms the N-ImageNet dataset. The performance of each method gets improved, when changing the pretraining dataset from N-ImageNet to E-TartanAir; ii) Our method consistently achieves the best performance under each setting, and the best performance of our method is achieved with the Swin-T/7 backbone and E-TartanAir dataset. Furthermore, our method gains performance improvements by at least 0.812% mIOU and 1.185% mACC scores over the state-of-the-art methods, with the same pre-training backbone and dataset. Note that ESViT exhibits numerical instability during pre-training, and the restart strategy from [50] is used for pre-training.

**Pre-training epochs.** We explore the impact of pre-training epochs, ranging from 100 to 800, and the results are presented in Fig. 5. Limited performance improvements in mIOU and mACC scores are observed after 300 epochs, prompting us to set the pre-training epoch number to 300.



**Fig. 5:** Comparison of the number of pre-training epochs.



**Fig. 6:** Sample results of patches belonging to different contexts on the E-TartanAir dataset. (a): input event images. (b): mined context labels (without enforcing the context-level similarity). (c): mined context labels (enforcing the context-level similarity). (d) and (e): blends of the event image with context labels from (b) and (c) for visualization purposes, respectively.

**Context-level similarity.** To check the effectiveness of our context-level similarity loss, we pre-train several networks without using it, varying our backbone network and pre-training dataset. Results in Tab. 7 reveal that a network pre-trained with  $\mathcal{L}_{\text{context}}$  consistently outperforms its counterpart pre-trained without using  $\mathcal{L}_{\text{context}}$ . For example, for networks pre-trained on the E-tartanAir dataset with the Swin-T/7 backbone, without using  $\mathcal{L}_{\text{context}}$  in pre-training, the mIOU/mACC scores are 55.556%/63.486%, which are lower than our best scores of 61.250%/69.620%. This justifies the effectiveness of the proposed context-level similarity loss.

Sample results of patches belonging to different contexts are given in Fig. 6. For example, in the 1<sup>st</sup> row, our method successfully mines contexts (tree, building, ground, and sky) in an event image, and groups patches with the same semantics.

**Number of contexts.** Our model utilizes  $K$  context embeddings by aggregating patch features. To study the impact of contexts, we train our model with a different number of contexts. Results in Fig. 7 indicate that the best performance is achieved with 8 contexts.

**Table 7:** Comparison of the performance of networks trained with and without using the proposed context-level similarity loss  $\mathcal{L}_{\text{context}}$ . Using  $\mathcal{L}_{\text{context}}$  consistently improves accuracies. ‘Pre. Dataset’ and ‘#Param’ respectively denote the pre-training dataset and the number of backbone parameters.

Pre. Dataset	Backbone	#Param	mIOU↑	mACC↑
$\mathcal{L}_{\text{patch}} + \mathcal{L}_{\text{image}}$ .				
N-ImageNet	ResNet50	23M	58.308	65.597
N-ImageNet	ViT-S/16	21M	53.706	61.328
N-ImageNet	Swin-T/7	28M	54.905	63.271
E-TartanAir	ResNet50	23M	58.687	66.171
E-TartanAir	ViT-S/16	21M	54.193	61.711
E-TartanAir	Swin-T/7	28M	55.556	63.486
$\mathcal{L}_{\text{patch}} + \mathcal{L}_{\text{context}} + \mathcal{L}_{\text{image}}$ .				
N-ImageNet	ResNet50	23M	60.243	69.195
N-ImageNet	ViT-S/16	21M	54.897	62.527
N-ImageNet	Swin-T/7	28M	56.654	65.250
E-TartanAir	ResNet50	23M	60.641	69.502
E-TartanAir	ViT-S/16	21M	55.729	64.771
E-TartanAir	Swin-T/7	28M	<b>61.250</b>	<b>69.620</b>

Increasing the number of contexts results in inferior performance. Due to the sparsity of event camera data, for large context numbers, many contexts aggregate features from event patches with little to no events. This results in noisy context embeddings, interferes with the training process, and hinders the network from learning discriminative event features.

#### Generalization ability of context-level similarity.

To further demonstrate the effectiveness of the proposed  $\mathcal{L}_{\text{context}}$ , we add  $\mathcal{L}_{\text{context}}$  to the objective function of the state-of-the-art event data pre-training method, ECDP [47]. The mIOU/mAcc scores of ECDP are increased by a large margin, having improvements from 52.517%/60.553% to 53.826%/61.008%. The improvements validate the generalization ability of  $\mathcal{L}_{\text{context}}$ .

**Limitation.** Although our self-supervised pre-trained network has achieved state-of-the-art performance across various dense prediction tasks, it necessitates task-specific fine-tuning to refine pre-trained network weights. However, we believe that our self-supervised learning exploration helps to learn task-agnostic pre-trained representations.

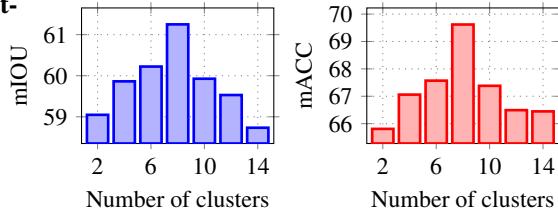
## 5 Conclusion and Broader Impact

We present a neural network trained for dense prediction tasks using an event camera. Our self-supervised learning method enforces three levels of similarity constraints: patch-level, context-level, and image-level. Our key insight is enforcing context similarity from event patch embeddings to pre-train our model. The proposed context-level similarity effectively addresses the sparsity problem of event data, resulting in state-of-the-art performance on semantic segmentation, optical flow, and depth estimation benchmarks. We believe that our dense pre-training techniques deserve a position in highly accurate event-based dense predictions.

**Broader Impact.** By aligning event data with paired RGB frames, our pre-training framework is promising to be extended to an event-vision-language foundation model. We hope it inspires future work.

## Acknowledgements

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



**Fig. 7:** Comparison of the number of contexts.

## References

1. Alonso, I., Murillo, A.C.: Ev-segnet: Semantic segmentation for event-based cameras. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 1624–1633. Computer Vision Foundation / IEEE (2019). <https://doi.org/10.1109/CVPRW.2019.00205>, [http://openaccess.thecvf.com/content\\_CVPRW\\_2019/html/EventVision/Alonso\\_EV-SegNet\\_Semantic\\_Segmentation\\_for\\_Event-Based\\_Cameras\\_CVPRW\\_2019\\_paper.html](http://openaccess.thecvf.com/content_CVPRW_2019/html/EventVision/Alonso_EV-SegNet_Semantic_Segmentation_for_Event-Based_Cameras_CVPRW_2019_paper.html)
2. Bai, Y., Chen, X., Kirillov, A., Yuille, A.L., Berg, A.C.: Point-level region contrast for object detection pre-training. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 16040–16049. IEEE (2022). <https://doi.org/10.1109/CVPR52688.2022.01559>, <https://doi.org/10.1109/CVPR52688.2022.01559>
3. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: BERT pre-training of image transformers. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022), <https://openreview.net/forum?id=p-BhZSz59o4>
4. Binas, J., Neil, D., Liu, S., Delbrück, T.: DDD17: end-to-end DAVIS driving dataset. CoRR **abs/1711.01458** (2017), <http://arxiv.org/abs/1711.01458>
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), <https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html>
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 9630–9640. IEEE (2021). <https://doi.org/10.1109/ICCV48922.2021.00951>, <https://doi.org/10.1109/ICCV48922.2021.00951>
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (2020), <http://proceedings.mlr.press/v119/chen20j.html>
8. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
9. Chen, X., He, K.: Exploring simple siamese representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 15750–15758. Computer Vision Foundation / IEEE (2021). <https://doi.org/10.1109/CVPR46437.2021.01549>, [https://openaccess.thecvf.com/content/CVPR2021/html/Chen\\_Exploring\\_Simple\\_Siamese\\_Representation\\_Learning\\_CVPR\\_2021\\_paper.html](https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html)
10. Chen\*, X., Xie\*, S., He, K.: An empirical study of training self-supervised vision transformers. arXiv preprint arXiv:2104.02057 (2021)
11. Cheng, W., Luo, H., Yang, W., Yu, L., Li, W.: Structure-aware network for lane marker extraction with dynamic vision sensor. CoRR **abs/2008.06204** (2020), <https://arxiv.org/abs/2008.06204>
12. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 2818–2829. IEEE (2023). <https://doi.org/10.1109/CVPR52729.2023.00276>, <https://doi.org/10.1109/CVPR52729.2023.00276>

13. Cuadrado, J., Rancon, U., Cottreau, B., Barranco, F., Masquelier, T.: Optical flow estimation from event-based cameras and spiking neural networks. *Frontiers in Neuroscience* **17** (05 2023). <https://doi.org/10.3389/fnins.2023.1160034>
14. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 20-25 June 2009, Miami, Florida, USA. pp. 248–255. IEEE Computer Society (2009). <https://doi.org/10.1109/CVPR.2009.5206848>, <https://doi.org/10.1109/CVPR.2009.5206848>
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net* (2021), <https://openreview.net/forum?id=YicbFdNTTy>
16. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. pp. 12873–12883. Computer Vision Foundation / IEEE (2021). <https://doi.org/10.1109/CVPR46437.2021.01268>, [https://openaccess.thecvf.com/content/CVPR2021/html/Esser\\_Taming\\_Transformers\\_for\\_High-Resolution\\_Image\\_Synthesis\\_CVPR\\_2021\\_paper.html](https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html)
17. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: EVA: exploring the limits of masked visual representation learning at scale. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. pp. 19358–19369. IEEE (2023). <https://doi.org/10.1109/CVPR52729.2023.01855>
18. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., Scaramuzza, D.: Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(1), 154–180 (2022). <https://doi.org/10.1109/TPAMI.2020.3008413>, <https://doi.org/10.1109/TPAMI.2020.3008413>
19. Gehrig, D., Rüegg, M., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics Autom. Lett.* **6**(2), 2822–2829 (2021). <https://doi.org/10.1109/LRA.2021.3060707>
20. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics Autom. Lett.* **6**(3), 4947–4954 (2021). <https://doi.org/10.1109/LRA.2021.3068942>
21. Gehrig, M., Millhäuser, M., Gehrig, D., Scaramuzza, D.: E-RAFT: dense optical flow from event cameras. In: *International Conference on 3D Vision, 3DV 2021, London, United Kingdom, December 1-3, 2021*. pp. 197–206. IEEE (2021). <https://doi.org/10.1109/3DV53792.2021.00030>
22. Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent - A new approach to self-supervised learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (2020), <https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html>
23. Hamaguchi, R., Furukawa, Y., Onishi, M., Sakurada, K.: Hierarchical neural memory network for low latency event processing. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. pp. 22867–

22876. IEEE (2023). <https://doi.org/10.1109/CVPR52729.2023.02190>, <https://doi.org/10.1109/CVPR52729.2023.02190>
24. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 15979–15988. IEEE (2022). <https://doi.org/10.1109/CVPR52688.2022.01553>, <https://doi.org/10.1109/CVPR52688.2022.01553>
  25. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 9726–9735. Computer Vision Foundation / IEEE (2020). <https://doi.org/10.1109/CVPR42600.2020.00975>, <https://doi.org/10.1109/CVPR42600.2020.00975>
  26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), <http://arxiv.org/abs/1512.03385>
  27. Kim, J., Bae, J., Park, G., Zhang, D., Kim, Y.M.: N-imagenet: Towards robust, fine-grained object recognition with event cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2146–2156 (October 2021)
  28. Li, C., Yang, J., Zhang, P., Gao, M., Xiao, B., Dai, X., Yuan, L., Gao, J.: Efficient self-supervised vision transformers for representation learning. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022), <https://openreview.net/forum?id=fVu3o-YUGQK>
  29. Li, W., Xie, J., Loy, C.C.: Correlational image modeling for self-supervised visual pre-training. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 15105–15115. IEEE (2023). <https://doi.org/10.1109/CVPR52729.2023.01450>, <https://doi.org/10.1109/CVPR52729.2023.01450>
  30. Li, Y., Huang, Z., Chen, S., Shi, X., Li, H., Bao, H., Cui, Z., Zhang, G.: Blinkflow: A dataset to push the limits of event-based optical flow estimation. CoRR **abs/2303.07716** (2023). <https://doi.org/10.48550/arXiv.2303.07716>, <https://doi.org/10.48550/arXiv.2303.07716>
  31. Liu, H., Chen, G., Qu, S., Zhang, Y., Li, Z., Knoll, A., Jiang, C.: Tma: Temporal motion aggregation for event-based optical flow. In: ICCV (2023)
  32. Menze, M., Heipke, C., Geiger, A.: Joint 3d estimation of vehicles and scene flow. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences **II-3/W5**, 427–434 (08 2015). <https://doi.org/10.5194/isprsannals-II-3-W5-427-2015>
  33. Oquab, M., Dariseti, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. CoRR **abs/2304.07193** (2023). <https://doi.org/10.48550/arXiv.2304.07193>, <https://doi.org/10.48550/arXiv.2304.07193>
  34. Orchard, G., Jayawant, A., Cohen, G., Thakor, N.V.: Converting static image datasets to spiking neuromorphic datasets using saccades. CoRR **abs/1507.07629** (2015), <http://arxiv.org/abs/1507.07629>
  35. Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F.: Beit v2: Masked image modeling with vector-quantized visual tokenizers. CoRR **abs/2208.06366** (2022). <https://doi.org/10.48550/arXiv.2208.06366>, <https://doi.org/10.48550/arXiv.2208.06366>
  36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), <http://proceedings.mlr.press/v139/radford21a.html>

37. Shiba, S., Aoki, Y., Gallego, G.: Secrets of event-based optical flow. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XVIII. Lecture Notes in Computer Science, vol. 13678, pp. 628–645. Springer (2022). [https://doi.org/10.1007/978-3-031-19797-0\\_36](https://doi.org/10.1007/978-3-031-19797-0_36), [https://doi.org/10.1007/978-3-031-19797-0\\_36](https://doi.org/10.1007/978-3-031-19797-0_36)
38. Sironi, A., Brambilla, M., Bourdis, N., Lagorce, X., Benosman, R.: HATS: histograms of averaged time surfaces for robust event-based object classification. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 1731–1740. Computer Vision Foundation / IEEE Computer Society (2018). <https://doi.org/10.1109/CVPR.2018.00186>, [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Sironi\\_HATS\\_Histograms\\_of\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Sironi_HATS_Histograms_of_CVPR_2018_paper.html)
39. Sun, Z., Messikommer, N., Gehrig, D., Scaramuzza, D.: ESS: learning event-based semantic segmentation from still images. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIV. Lecture Notes in Computer Science, vol. 13694, pp. 341–357. Springer (2022). [https://doi.org/10.1007/978-3-031-19830-4\\_20](https://doi.org/10.1007/978-3-031-19830-4_20), [https://doi.org/10.1007/978-3-031-19830-4\\_20](https://doi.org/10.1007/978-3-031-19830-4_20)
40. Wan, Z., Dai, Y., Mao, Y.: Learning dense and continuous optical flow from an event camera. *IEEE Trans. Image Process.* **31**, 7237–7251 (2022). <https://doi.org/10.1109/TIP.2022.3220938>, <https://doi.org/10.1109/TIP.2022.3220938>
41. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.A.: Tartanair: A dataset to push the limits of visual SLAM. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2020, Las Vegas, NV, USA, October 24, 2020 - January 24, 2021. pp. 4909–4916. IEEE (2020). <https://doi.org/10.1109/IROS45743.2020.9341801>, <https://doi.org/10.1109/IROS45743.2020.9341801>
42. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 3024–3033. Computer Vision Foundation / IEEE (2021). <https://doi.org/10.1109/CVPR46437.2021.00304>, [http://openaccess.thecvf.com/content/CVPR2021/html/Wang\\_Dense\\_Contrastive\\_Learning\\_for\\_Self-Supervised\\_Visual\\_Pre-Training\\_CVPR\\_2021\\_paper.html](http://openaccess.thecvf.com/content/CVPR2021/html/Wang_Dense_Contrastive_Learning_for_Self-Supervised_Visual_Pre-Training_CVPR_2021_paper.html)
43. Weikersdorfer, D., Adrian, D.B., Cremers, D., Conradt, J.: Event-based 3d SLAM with a depth-augmented dynamic vision sensor. In: 2014 IEEE International Conference on Robotics and Automation, ICRA 2014, Hong Kong, China, May 31 - June 7, 2014. pp. 359–364. IEEE (2014). <https://doi.org/10.1109/ICRA.2014.6906882>, <https://doi.org/10.1109/ICRA.2014.6906882>
44. Wu, Y., Paredes-Vallés, F., de Croon, G.C.H.E.: Lightweight event-based optical flow estimation via iterative deblurring. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA'24) (May 2024), to Appear
45. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 3733–3742. Computer Vision Foundation / IEEE Computer Society (2018). <https://doi.org/10.1109/CVPR.2018.00393>, [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Wu\\_Unsupervised\\_Feature\\_Learning\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Wu_Unsupervised_Feature_Learning_CVPR_2018_paper.html)
46. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: a simple framework for masked image modeling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 9643–9653. IEEE (2022). <https://doi.org/10.1109/CVPR52688.2022.00943>, <https://doi.org/10.1109/CVPR52688.2022.00943>

47. Yang, Y., Pan, L., Liu, L.: Event camera data pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10699–10709 (October 2023)
48. Yun, S., Lee, H., Kim, J., Shin, J.: Patch-level representation learning for self-supervised vision transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 8344–8353. IEEE (2022). <https://doi.org/10.1109/CVPR52688.2022.00817>, <https://doi.org/10.1109/CVPR52688.2022.00817>
49. Zhang, D., Ding, Q., Duan, P., Zhou, C., Shi, B.: Data association between event streams and intensity frames under diverse baselines. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VII. Lecture Notes in Computer Science, vol. 13667, pp. 72–90. Springer (2022). [https://doi.org/10.1007/978-3-031-20071-7\\_5](https://doi.org/10.1007/978-3-031-20071-7_5)
50. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M.T., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L.: OPT: open pre-trained transformer language models. CoRR **abs/2205.01068** (2022). <https://doi.org/10.48550/ARXIV.2205.01068>, <https://doi.org/10.48550/arXiv.2205.01068>
51. Zhou, J., Zheng, X., Lyu, Y., Wang, L.: E-CLIP: towards label-efficient event-based open-world understanding by CLIP. CoRR **abs/2308.03135** (2023). <https://doi.org/10.48550/arXiv.2308.03135>
52. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A.L., Kong, T.: Image BERT pre-training with online tokenizer. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022), <https://openreview.net/forum?id=ydopy-e6Dg>
53. Zhu, A.Z., Thakur, D., Özaslan, T., Pf frommer, B., Kumar, V., Daniilidis, K.: The multi vehicle stereo event camera dataset: An event camera dataset for 3d perception. CoRR **abs/1801.10202** (2018), <http://arxiv.org/abs/1801.10202>
54. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. CoRR **abs/1812.08156** (2018), <http://arxiv.org/abs/1812.08156>