Efficient Event Camera Data Pretraining with Adaptive Prompt Fusion

Quanmin Liang[1,2*] Qiang Li[4*] Shuai Liu[1] Xinzi Cao[1,2] Jinyi Lu[1,2]
Feidiao Yang[2] Wei Zhang[2†] Kai Huang[1†] Yonghong Tian[2,3]
[1] School of Computer Science and Engineering, Sun Yat-Sen University
[2] Department of Intelligent Computing, Pengcheng Laboratory
[3] School of Computer Science, Peking University [4] Xpeng Motors Technology Co Ltd
{liangqm5@mail2,liqiang27@mail2,huangk36@mail}.sysu.edu.cn, zhangwei1213052@126.com

Abstract

*Applying pretraining-finetuning paradigm to event cameras presents significant challenges due to the scarcity of large-scale event datasets and the inherently sparse nature of event data, which increases the risk of overfitting during extensive pretraining. In this paper, we explore the transfer of pre-trained image knowledge to the domain of event cameras to address this challenge. The key to our approach lies in adapting event data representations to align with image pretrained models while simultaneously integrating spatiotemporal information and mitigating data sparsity. To achieve this, we propose a lightweight SpatioTemporal information fusion Prompting (**STP**) method, which progressively fuses the spatiotemporal characteristics of event data through a dynamic perception module with multi-scale spatiotemporal receptive fields, enabling compatibility with image pretrained models. STP enhances event data representation by capturing local information within a large receptive field and performing global information exchange along the temporal dimension. This strategy effectively reduces sparse regions in event data while refining fine-grained details, all while preserving its inherent spatiotemporal structure. Our method significantly outperforms previous state-of-the-art approaches across classification, semantic segmentation, and optical flow estimation tasks. For instance, it achieves a top-1 accuracy of **68.87% (+4.04%)** on N-ImageNet with only 1/10 of the pretraining parameters and 1/3 of the training epochs. Our code is available at https://github.com/Lqm26/STP.*

1. Introduction

Event cameras are dynamic vision sensors inspired by the perceptual mechanism of the human retinas [39, 57, 63]. They asynchronously capture the event stream by comparing the intensity changes of each pixel [6, 18, 39, 63]. This triggering mechanism enables event cameras to efficiently record information in high-dynamic-range (HDR, 120dB) or high-speed motion scenes, while offering advantages such as low power consumption and low redundancy [39]. Currently, event cameras are widely used in novel computer vision and robotics tasks, including video interpolation [19, 60, 64, 78], image or video reconstruction [38, 48, 51, 56], optical flow estimation [34, 84], depth estimation [17, 86], super-resolution [16, 29, 37], and SLAM [31, 47, 66].

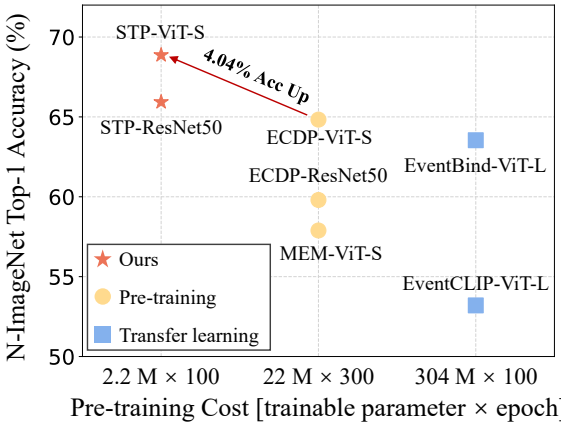

Figure 1. Comparison of our method with existing pre-training and transfer learning approaches on the N-ImageNet dataset [32].

However, due to the high cost of acquiring event camera data and the difficulty of labeling, there is still a lack of pre-trained models based on large-scale event camera datasets. This has hindered the adoption of the pretraining-finetuning paradigm for event-based vision tasks and limited the development of corresponding deep learning methods and models. Given the emergence of large-scale RGB image datasets (e.g., ImageNet-21k [13], JFT-300M [59]) and the development of pretrained models [15, 23, 25, 53] based on these datasets, researchers have attempted to use transfer learning

*Equal Contribution
†Corresponding Author

[27, 46, 61] or knowledge distillation [69] to transfer the knowledge from RGB image-trained models to event-based downstream tasks. On the other hand, some researchers have attempted to use pretrained image models or VLMs [53] for object recognition [73, 81–83], employing few-shot or unsupervised methods. However, these efforts have been limited to single tasks and have not extended to a broader range of downstream applications, preventing them from serving as more versatile pretraining models.

To address the lack of event-based pretrained models, some researchers have attempted pretraining on the N-ImageNet dataset [32]. They have demonstrated that the pretraining-finetuning paradigm is effective on event camera data as well [33, 76, 77]. However, these approaches fail to fully consider the unique characteristics of event data, such as spatiotemporal features, polarity, and sparsity. Moreover, the largest available event camera dataset is derived from ImageNet-1K [13], which remains significantly smaller compared to large-scale RGB image datasets. As a result, pretrained models for event data may suffer from overfitting due to an imbalance between the large number of trainable parameters and the limited amount of available training data.

In summary, **pretraining on event camera data faces three main challenges**: (i). The sparsity of event data, which can lead to model overfitting. (ii). The challenge of integrating multiple attributes of event data. (iii). The requirement for sufficient prior knowledge to mitigate the imbalance between a large number of trainable parameters and limited data.

To address these challenges, we propose a method that transfers knowledge from the image domain to the event camera domain through carefully designed prompting mechanisms. Specifically, we introduce a Spatiotemporal Information Fusion Module tailored for event data within the prompting process. This module leverages overlapping convolutions with large receptive fields and a Temporal Transformer to achieve multi-scale spatiotemporal dynamic perception. By progressively integrating the temporal and spatial information of event streams, our approach effectively reduces sparse regions in event data and generates a Dense Event Stack (DES). The DES is then fed into a pretrained image model with frozen weights to extract high-level features for classification. This end-to-end optimization framework enables the transfer of prior knowledge from the image domain to guide the training of the STP, facilitating the fusion of spatiotemporal information and enhancing knowledge transfer. Finally, STP, in combination with image pretrained models, forms a pretrained model for event camera data. This model can be fine-tuned on downstream tasks, leveraging the wealth of image pre-trained knowledge as prior information. This significantly reducing the imbalance between training parameters and limited event data.

In summary, our work makes the following contributions:

- We propose an efficient, lightweight, and versatile pretraining paradigm for event camera data. Our approach employs adaptive prompt-based fusion to transfer image-pretrained knowledge to the event camera domain, enabling efficient pretraining across multiple tasks.
- We designed a Spatiotemporal Information Fusion Prompting Module tailored to the characteristics of event data, which progressively integrates the spatiotemporal information of events through multi-scale spatiotemporal dynamic perception, reducing the sparse regions in event data.
- Our method achieves SOTA performance across downstream tasks such as classification, semantic segmentation, and optical flow estimation. For instance, we achieve a top-1 accuracy of **68.87% (+4.04%)** on the N-ImageNet.

2. Related Work

2.1. Visual Pre-training

With the rapid development of deep learning and computer vision, pretrained large models have become an important topic and research method [24, 26, 54], driven by the continuous evolution of visual models. From the perspective of training, these methods mainly include supervised pre-training on large-scale datasets [8, 12, 15, 79], weakly supervised pre-training requiring less data [4, 52, 55, 75, 80], and unsupervised pre-training that exploits intrinsic features of data for learning without relying on any labels [3, 9, 10, 22, 25]. These models can efficiently transfer knowledge to downstream tasks through tuning.

In contrast to the rapid development of image-based pretraining, event-based pre-training is still in its early stages. There are two main challenges in pre-training with event camera data: first, the difficulty in acquiring event stream data and the lack of large-scale datasets; second, the sparsity of event stream data, which easily leads to overfitting or training collapse during extensive training. Previous pretraining methods have been primarily relied on self-supervised learning. Yang et al. [76], were the first to propose a contrastive learning-based method for large-scale event camera data pretraining. They later introduced a self-supervised method for dense prediction tasks, such as depth estimation, which automatically mines contextual similarities between patches [77]. Klenk et al. [33], drew inspiration from VQVAE's discrete encoding [65] and BERT's masked reconstruction [14] to propose Masked Event Modeling (MEM). Huang et al. [28] proposed a self-supervised learning method based on voxel data, enabling fast convergence with a small amount of pretraining data.

2.2. Prompt Tuning

Prompt tuning is an important paradigm that leverages pretrained large models [35, 40]. As a lightweight tuning method, its principle is to adapt downstream tasks to the

original training task at minimal cost, thereby utilizing the knowledge embedded in pretrained models to address problems. Prompting was initially introduced in natural language processing (NLP) [40], where additional tokens are added to token sequences to help the pretrained model better "understand" the task [35, 36]. Initially, the values of prompt engineering were heuristically selected [7]. Subsequently, prompt methods based on learnable parameters gradually became mainstream due to their efficiency and flexibility [35, 36, 41, 67]. Due to its simplicity and effectiveness, prompt tuning has also been applied to some visual tasks such as image classification [2, 30], segmentation [49], and 3D point clouds [62, 72, 87].

Inspired by these studies, Wu et al. explored using CLIP [53] for event image recognition [73]. Zhou et al. proposed a language-guided approach for event action recognition [83]. They extended CLIP by incorporating an Event Encoder to align text, image, and event data [82]. Zheng et al. took an unsupervised approach to extract pretrained image knowledge for object recognition [81]. However, each of these methods is tailored to a single recognition task, limiting their generalizability to other tasks.

3. Method

3.1. Event Preparation

Integrating event stream data with deep learning typically requires flattening the event stream into a 2D representation. Common approaches include the Event Count Image (ECM) [44, 84], Voxel grid [86], and Event Spike Tensor (EST) [20]. However, for event camera data pre-training, we aim for an input representation that retains as many of the event stream's intrinsic features and information as possible, including temporal details, spatial distribution, and event polarity, while avoiding the addition of extraneous information that could hinder generalization.

Inspired by [88], we propose a Temporal Event Count Map (TECM), which builds on the ECM by incorporating temporal information, effectively resolving its limitation of discarding time information. We illustrate ECM and TECM in Fig. 2. Specifically, following the approach of the Voxel grid, we divide the event stream into T temporal segments. For each segment of the event stream $\mathcal{E}_T = \{e_k\}_{k=1}^{n}$, where n is the number of events, each event e_k is represented by a tuple (x_i, y_i, t_i, p_i), where x_i and y_i denote the pixel coordinates, t_i represents the timestamp of the event, and $p_i = \pm 1$ represents the polarity. As event cameras asynchronously report changes in pixel intensity, the output is a series of independent events [6, 18]. To obtain a 2D image structure with visible edges, we accumulate events of different polarities separately onto a 2D plane, resulting in a $ECM \in \mathbb{R}^{2\times H\times W}$. As illustrated in Fig. 3, by concatenating these T ECM, we obtain a temporal ECM representation

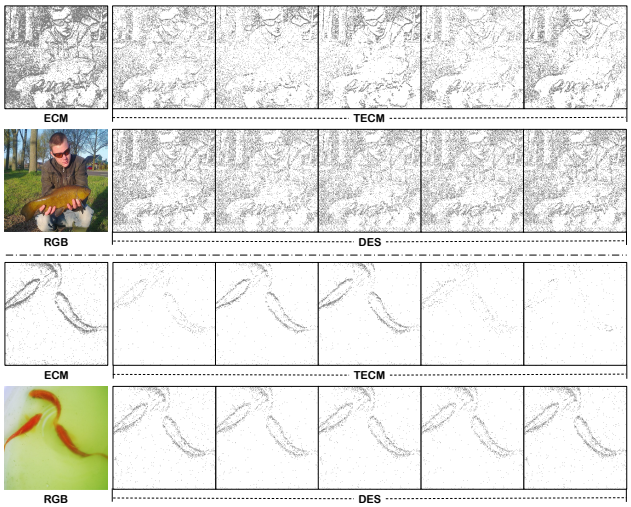

Figure 2. The representations of event data (ECM and TECM) and their corresponding RGB images are shown, with DES generated from TECM through STP. For ease of visualization, we overlay positive and negative events.

(TECM) of the event stream. This representation method retains the complete spatial structure and effective temporal information of the event stream without introducing additional information or constraints. As a result, it can be effectively transferred to a variety of downstream tasks.

3.2. SpatioTemporal Prompting

Our STP follows the principles of prompt-tuning, aiming to maintain minimal computational and parameter overhead. The proposed STP is a dynamic perception fusion module based on multi-scale spatiotemporal receptive fields, achieving efficient spatiotemporal feature extraction through the progressive collaboration of Overlap Patch Embedding and Temporal Transformer. Specifically: (i). Adjustable large-kernel convolutions are used to construct multi-level patch embeddings, which exponentially expand the local receptive field and progressively fill sparse regions. (ii). The Temporal Transformer employs a self-attention mechanism and a local temporal aggregation (LTA) to achieve both global and local cross-frame dynamic alignment. (iii). Through a progressive design, efficient spatiotemporal feature fusion is achieved at multiple scales, from local to global.

3.2.1. Overlap Patch Embedding

In the Overlap Patch Embedding, the kernel size of the convolution must be larger than the stride to utilize information from neighboring patches and fill in sparse regions within the current patch. Specifically, given an input event image with dimensions $T \times H \times W \times C$, where T, H, W and C represent the temporal dimension of the event stream, image height, image width, and channel, respectively. We apply a convolution layer with a stride of P and a kernel

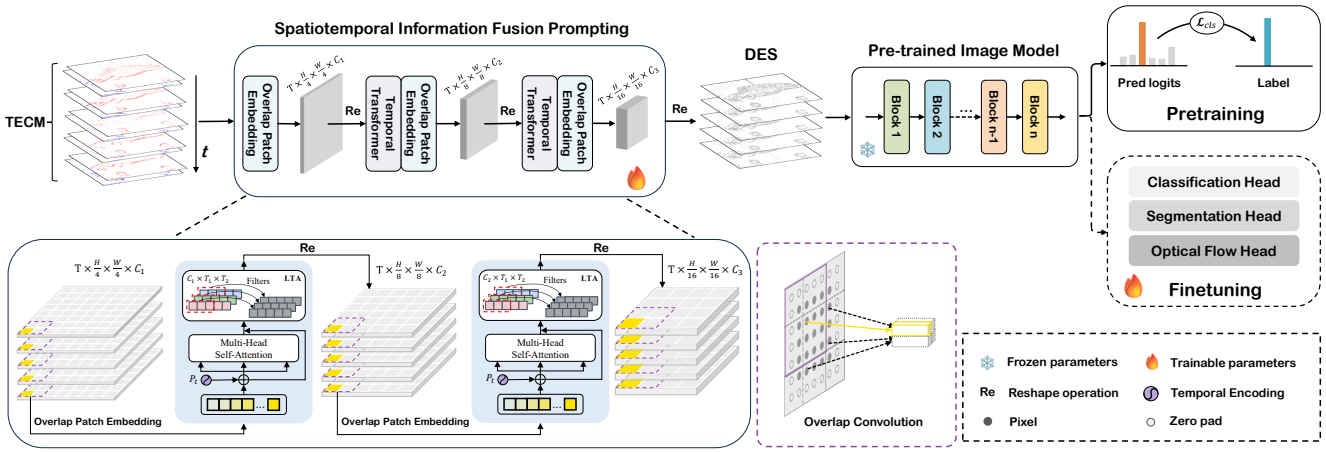

Figure 3. **Pipeline of STP**. First, the event stream data is converted into TECM and concatenated. Then, STP fuses the TECM to produce a Dense Event Stack (DES). STP consists of two key components: Overlap Patch Embedding and the Temporal Transformer, where Overlap Patch Embedding is primarily implemented using Overlap Convolution. Finally, DES is fed into a frozen pretrained image model for classification, with the classification loss $\mathcal{L}_{cls}$ guiding the training of STP.

size K, where $K > P$. The size of the output feature map is $T \times \frac{H}{P} \times \frac{W}{P} \times CP^2$. Next, we apply a 3×3 convolution to extract finer local details, while keeping the feature map size unchanged. Finally, the feature map is reshaped to size $\left(\frac{H}{P} \times \frac{W}{P}\right) \times T \times CP^2$ and fed into the Temporal Transformer, which fuses temporal information within each patch. The Overlap Patch Embedding is applied three times within the STP, with the kernel size $\{k_1, k_2, k_3\}$ being a set of hyperparameters. The strides are set to $\{4, 2, 2\}$. By progressively increasing the receptive field, each patch is enriched with information from neighboring patches, thereby enhancing its embedding representation and mitigating the risk of overfitting due to sparsity.

Analysis. Compared to the standard ViT Patch Embedding [15], our Overlap Patch Embedding significantly expands the receptive field of each patch. The receptive field size for the first patch can be computed using the following formula:

$$R_l = k_1 - p_1 + \sum_{i=1}^{l}(k_i - 1) \times \prod_{j=1}^{i-1} s_j \tag{1}$$

where R_l denotes the receptive field size at layer l, k_i represents the kernel size of the i-th convolutional layer, s_j is the stride of the j-th layer, and p denotes padding. In a standard Patch Embedding, each patch has a receptive field of only 16×16. However, with Overlap Patch Embedding and convolutional kernel sizes of $\{8, 6, 6\}$, the receptive field expands to 66×66, an increase of over 16 times. Moreover, we adopt a progressive embedding strategy to gradually enlarge the receptive field. This, combined with the Temporal Transformer, enhances global spatiotemporal information perception while maintaining sensitivity to local details, thereby reducing the risk of overfitting [23].

3.2.2. Temporal Transformer

As shown in Fig. 3, the Temporal Transformer primarily follows the Transformer block design of ViT [15]. For each patch, we fuse and enrich contextual information along the temporal dimension T. The Overlap Patch Embedding further expands each patch's receptive field, promoting spatiotemporal information fusion. Additionally, we introduce a local temporal aggregation (LTA) layer between two linear layers, enhancing local information exchange among neighboring events. Thus, our Temporal Transformer can be represented by the following equation:

$$\boldsymbol{z}_0 = \oplus_{t=1}^{T} \phi_{\text{proj}}(\boldsymbol{x}_{\text{TECM}}^{t}) + \Psi_{\text{tem}}(T) \in \mathbb{R}^{N_p \times T \times CP^2} \tag{2}$$

where $N_p = \frac{HW}{P^2}$ denotes the number of spatial positions, $\boldsymbol{x}$ denotes each temporal token, ϕ_{proj} refers the patch embedding operator, Ψ_{tem} denotes the learnable temporal encoding, and $\oplus$ refers concatenation.

$$\boldsymbol{z}_l^{'} = MHA\left(LN\left(\boldsymbol{z}_{l-1}\right)\right) + \boldsymbol{z}_{l-1} \tag{3}$$

MHA and LN stand for Multi-Head Attention and LayerNorm [15], respectively. The Temporal Transformer computes attention as $\text{Attention}(Q \in \mathbb{R}^{N_p \times T \times P^2}, K^{\mathrm{T}} \in \mathbb{R}^{N_p \times P^2 \times T}, V \in \mathbb{R}^{N_p \times T \times P^2})$. This attention operates on the temporal dimension T of each patch.

$$\boldsymbol{z}_l = Linear\left(LTA\left(Linear\left(LN\left(\boldsymbol{z}_l^{'}\right)\right)\right)\right) + \boldsymbol{z}_l^{'} \tag{4}$$

$$LTA(c,t) = \sum_{\tau=-1}^{+1} w_\tau \cdot F(c, t+\tau) \tag{5}$$

where w_τ denotes the learnable aggregation weights, F represents the feature map, and c indexes the channel dimension.

Method	Backbone	Pr. Params	Pr. Epoch	N-ImageNet		N-Caltech101	N-Cars	CIF10
				acc@1	acc@5			
	Training from scratch							
EST [20]	-	21M	-	48.93	-	68.12	90.80	62.57
ViT [15]	ViT-S/16	22.1M	-	46.70	69.89	55.63	89.14	52.45
ViT [15]	ViT-B/16	86.6M	-	51.23	74.50	67.11	93.09	55.15
ResNet [23]	ResNet50	25.6M	-	50.07	74.83	62.69	91.20	56.65
	Transfer learning from models pretrained on ImageNet [13]							
N-ImageNet [32]	-	-	-	-	-	80.88	91.48	70.36
ViT [15]	ViT-S/16	22.1M	300	60.48	83.02	85.02	96.76	76.10
ViT [15]	ViT-B/16	86.6M	300	62.98	84.75	86.45	97.56	77.45
ResNet [23]	ResNet50	25.6M	90	57.37	80.93	86.51	97.61	73.40
	Incorporating image pretrained models							
EventCLIP [73]	ViT-L/14	304.3M	100	53.20	-	93.57	90.34	-
EventBind [82]	ViT-L/14	304.3M	100	63.54	-	95.29	-	-
EventDance [81]	ResNet50	25.6M	-	-	-	92.30	-	85.69
Ours (w/o finetuning)	ViT-S/16	2.2M	100	66.01	88.14	-	-	-
	Pretraining on N-ImageNet [32] + Finetuning							
EST [20]	-	21M	-	-	-	86.81	94.73	73.72
ECDP [76]	ViT-S/16	22.1M	300	<u>64.83</u>	<u>86.30</u>	87.66	<u>97.93</u>	78.00
MEM [33]	dVAE+VIT	23.1M	125	57.89	-	<u>90.10</u>	93.27	-
DMM [28]	-	<u>13.5M</u>	700	-	-	88.00	97.10	<u>78.60</u>
Ours	ViT-S/16	**2.2M**	100	**68.87**	**89.65**	**94.74**	**98.86**	**88.76**

Table 1. Comparison of object recognition accuracy, trainable parameters during the pretraining stage (Pr. Params), and pre-training epochs (Pr. Epoch) on N-ImageNet, N-Caltech101, N-Cars, and CIFAR-10-DVS datasets. Top-1 (acc@1) and top-5 (acc@5) accuracy are shown on N-ImageNet, while only top-1 accuracy is reported on small-scale datasets. '-' indicates either the result is not reported or not supported by the method. **Bold** and <u>underline</u> indicate the best and second-best results.

After passing through the first linear layer, the dimension of $\boldsymbol{z}$ becomes $N_p \times T \times CP^2R$, which is then reshaped into $N_p \times CP^2R \times T \times 1$, where R denotes MLP ratio. We then apply a LTA (Eq. 5) for local temporal aggregation while keeping the shape unchanged. Finally, $\boldsymbol{z}$ is reshaped back and another linear layer is applied to adjust the dimension to $N_p \times T \times CP^2$. ***Analysis***. In the Temporal Transformer, our key improvement is the introduction of a local temporal aggregation (LTA) layer, motivated by two main reasons. First, to maintain a lightweight design, our STP model uses only two attention layers, far fewer than the 12+ used in standard ViTs. While this may limit temporal modeling capacity, we address this with the LTA module, which enhances inter-frame correlations through localized feature interactions, ensuring effective temporal fusion with minimal parameter overhead. Second, in event streams, events within a given time segment are most correlated with those in adjacent segments. LTA strengthens local information interactions while preserving global temporal information exchange.

3.3. Training Objectives

Pretraining. After processing through STP, we reshape the features $\boldsymbol{z}$ into the $DES \in \mathbb{R}^{T\times H\times W}$ and input it into the pretrained image model. In the pretraining phase, we utilize ViT [15] pretrained on ImageNet [13] as the backbone of our method and freeze its weights. Additionally, we replace and train the patch embedding layer in ViT to better adapt to event data. The DES from the STP are input into the pretrained model, resulting in a class token feature $f_{cls} \in \mathbb{R}^{1\times C_t}$, where C_t represents the token feature dimension. Finally, target classification is performed based on the class token, and the knowledge from the image pretrained model is transferred to STP through the Cross-Entropy loss $\mathcal{L}_{cls}$.
Finetuning. During the finetuning stage, both the weights of the STP and ViT need to be trained to better adapt to downstream tasks. The classification head can be adjusted based on the specific types of data in the downstream tasks, while the other structures remain unchanged.

4. Experiments

4.1. Dataset and Experimental Setup

Pre-training Dataset. We utilize the N-ImageNet [32] for pre-training. N-ImageNet is the largest event camera classification dataset to date, generated by capturing ImageNet-1K images displayed on a screen using a moving event camera. To ensure compatibility with the pretrained model, we resize

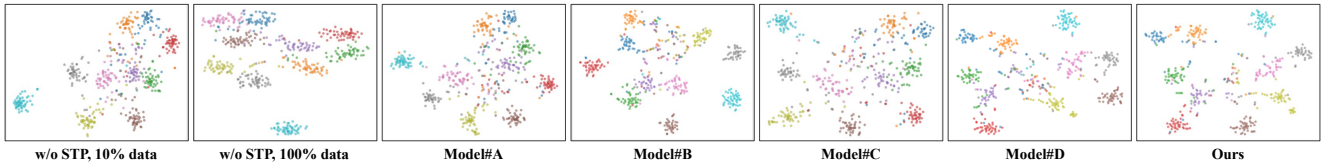

Figure 4. t-SNE visualization of 10 selected classes from N-ImageNet [32]. Here, **w/o STP** refers to training ViT from scratch without the STP module, while **10% data** indicates training using only 10% of the N-ImageNet training set. **Model#A-D** represent different variants of STP used in the ablation study (Sec. 4.3).

the event data to a resolution of 224×224.

4.2. Object Classification

In this task, we primarily compare our approach with previous methods pretrained on N-ImageNet: ECDP [76] and MEM [33]. Additionally, Huang et al. [28] proposed a data-efficient method, DMM. However, due to its requirement for longer event stream durations, DMM could only be pretrained on the N-Caltech101 dataset. Following the approach of ECDP [76], we include the results of training from scratch, and transfer learning from pretrained models, for comparison. Additionally, we incorporate some classification models based on image pretrained models into the comparison [73, 81, 82]. We conduct performance comparisons and analyses on both the large-scale dataset N-ImageNet and small-scale datasets N-Caltech101 [50], N-Cars [58], and CIFAR-10-DVS [11].

Results on N-ImageNet. As shown in Tab. 1, our method achieved a top-1 accuracy of 68.87%, surpassing the previous best result (64.83%) by **4.04%**. Remarkably, even at the end of the pretraining phase, our accuracy reached 66.01% (Tab. 1), surpassing the prior best result. This demonstrates that our approach effectively leverages pretrained knowledge from the image domain to enhance event image classification performance. In contrast to other methods that require pretraining on large-parameter backbones such as VIT and ResNet, our STP has only 2.2M trainable parameters and requires training for only 100 epochs, effectively reducing the demand for training resources.

Results on small-scale datasets. As shown in Tab. 1, our method achieved top-1 accuracies of 94.74%, 98.86%, and 88.76% on the N-Caltech101, N-Cars, and CIFAR-10-DVS datasets, respectively. Compared with the previous SOTA methods, our approach improved by **4.64%**, **0.93%**, and **10.16%**, respectively. This demonstrates the efficient transfer of knowledge from pretrained image models to the event camera domain.

Visualization Analysis. As shown in Fig. 2, we visualize TECM and the generated DES. To facilitate a more direct comparison, we only display the DES events within the ECM, as it includes the full spatial distribution of the original event data. It can be observed that STP effectively reduces sparse regions while enhancing fine-grained details based on context-aware event information.

Variants	OPE		LTA	Pr. Acc	Ft. Acc
	Overlap	Conv			
Model#A	✘	✔	✔	64.86	67.04
Model#B	✔	✘	✔	65.77	67.36
Model#C	✘	✘	✔	59.30	63.16
Model#D	✔	✔	✘	65.44	68.12
STP	✔	✔	✔	**66.01**	**68.87**

Table 2. Variants of the STP. **Pr. Acc** and **Ft. Acc** represent the Top-1 accuracy after pretraining and fine-tuning, respectively.

4.3. Ablation Studies

To validate the effectiveness of our proposed framework and STP, we conducted extensive ablation studies on the N-ImageNet classification task. These studies includes different variants of STP, various prompting models, different event data input methods, different pretraining model backbones, as well as model hyperparameter settings (see **Supplement** for details).

(1) Variants of the STP. Our proposed STP method consists of two key components: Overlap Patch Embedding (OPE) and the Temporal Transformer. We conducted ablation studies to examine different variations of these components: 1) **Model#A**: We removed the Overlap window mechanism and used independent patch division similar to ViT with position embeddings. 2) **Model#B**: We eliminated the fine-grained information extraction layer ($Conv$) following the Patch Embedding. 3) **Model#C**: Both the Overlap window and the fine-grained information extraction layer were removed. 4) **Model#D**: Removed the local temporal aggregation (LTA) layer from the Temporal Transformer.

Quantitative Analysis. As shown in Tab. 2, compared to **Model#A-C**, our progressive multi-scale spatiotemporal overlapping receptive field is crucial for information fusion in STP, effectively reducing sparse regions and improving performance. The comparison with **Model#D** further demonstrates that LTA contributes to local information aggregation across frames. Notably, adding LTA only increases parameters by 6.4K (**+0.75 Acc**), highlighting its efficiency.

Visualization Analysis. (i). In Fig. 4, we apply t-SNE di-

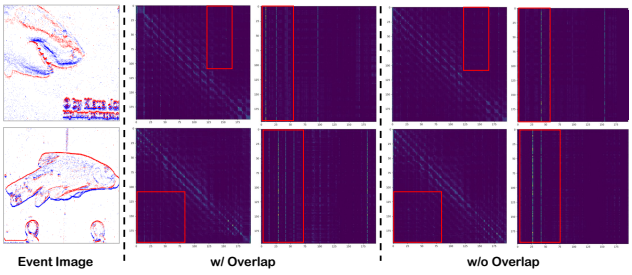

Figure 5. **Visualization of event images and their corresponding attention matrices.** For models w/ Overlap and w/o Overlap, from left to right are attention weights from the 6*th*, and 12*th* layers. The comparison shows that the attention matrix of model w/ Overlap is more uniformly distributed (highlighted in the red box).

mensionality reduction to visualize and compare different STP variants. To further analyze the effectiveness of the STP module and the impact of pretraining dataset size on model performance, we introduce two additional comparisons: training from scratch without STP, using either the full N-ImageNet training set or only 10% of the data. It can be observed that without STP, inter-class separation is smaller, and intra-class distributions are more scattered, which cannot be effectively resolved by simply increasing the dataset size. In contrast, comparisons with **Model#A-D** show that the multi-scale spatiotemporal overlapping receptive field effectively captures local information and reduces sparse regions, leading to clearer inter-class separation and tighter intra-class clustering. Additionally, LTA further enhances this effect. (ii). We also investigate the impact of the overlapping mechanism on internal attention patterns. As shown in Fig. 5, we feed DES, which incorporates fused spatiotemporal information, into the image pretrained model and visualize the attention weights at the 6*th* and 12*th* layers. The results show that with the overlapping mechanism, the attention weight matrix is more evenly distributed, whereas without it, the weights are more concentrated. This further demonstrates that the overlapping mechanism effectively fills sparse regions in event data, leading to a more comprehensive and robust representation.

(2) Prompting Models. To validate the effectiveness of our proposed STP, we replaced the STP module with other methods. One approach followed previous practices by reconstructing the event stream into images [61] and adapting them to a pretrained image model. We used the E2VID [56] method as the reconstruction model. Additionally, we designed a vanilla for event data prompting (STP-vanilla), where we first encode the TECM using a 16×16 patch embedding, then pass it through two transformer blocks identical to those in ViT [15]. As shown in Tab. 3, STP can more effectively fuses and extracts event information through progressive multi-scale spatiotemporal perception.

(3) Event Data Representation Methods. We compared different event stream representation methods, including ECM [44], Voxel grid [86], Polarity Voxel Grid [88], and EST [20], against our proposed TECM. We replicate ECM T times and concatenate them to match the dimensions of TECM. As shown in Tab. 4, Voxel Grid discards event polarity and disrupts the spatial structure, leading to performance degradation. EST introduces additional constraints on the event stream, reducing generalization ability. Meanwhile, ECM lacks temporal information and fails to fully represent the event stream. The results demonstrate that TECM effectively captures event stream information, improving performance in classification tasks.

Prompting	#Params	FLOPs (G)	Pr. Acc	Ft. Acc
E2VID [56]	4.5 M	11.91	58.24	64.93
STP-vanilla	4.1 M	4.26	61.94	66.89
STP	**2.2 M**	**3.55**	**66.01**	**68.87**

Table 3. Ablation Study on Prompting Model.

Representation	Dimensions	Acc@ 1 (Δ)
ECM [84]	$2 \times H \times W$	67.82 (+2.99)
Voxid Grid [86]	$T \times H \times W$	58.26 (−6.57)
Polarity Voxel Grid [88]	$2 \times T \times H \times W$	66.86 (+2.03)
EST [20]	$2 \times T \times H \times W$	68.07 (+3.24)
TECM	$2 \times T \times H \times W$	**68.87** (+4.04)

Table 4. Ablation Study on Event Data Input Methods. Δ represents the relative top-1 accuracy compared to ECDP [76].

Pr. Arch	#Params	Pr. Acc	Ft. Acc	FLOPs (G)
ResNet50 [23]	25.6M	60.70	65.93	**7.34**
ConvNeXt_T [43]	28.6M	67.09	71.90	7.69
Swin_T [42]	28.3M	64.18	68.34	7.65
ViT-S/16 [15]	**22.1M**	66.01	68.87	7.84
ViT-B/16 [15]	86.6M	73.14	75.88	20.84
ViT-L/16 [15]	304.3M	**75.51**	**78.06**	64.84
BEVT [70]	88M	72.06	74.46	286

Table 5. Ablation Study on the Backbone of Pretrained Models.

(4) Ablation Study on Pretraining Models. Our method initially uses VIT-S/16 pretrained on ImageNet. For comparison, we replaced it with three image classification pretrained models with comparable parameter counts: ResNet50 [23], Swin-T [42], and ConvNeXt-T [43]. We also explored scaling VIT parameters, using VIT-B/16 and VIT-L/16, and further compared our approach with the video-based pretrained model BEVT [70]. As shown in Tab. 5, our method achieves superior results across various image pre-training models, surpassing the previous SOTA methods. Notably, our approach achieved a top-1 accuracy of 78.06% with VIT-L/16, marking the first instance of surpassing 75% accuracy on the N-ImageNet. This further demonstrates the strong adaptability of our method to various pretrained models.

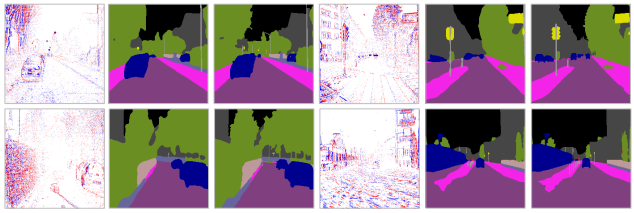

Figure 6. Examples of semantic segmentation on the DSEC dataset. Columns 1/4 show event images, columns 2/5 show segmentation results, and columns 3/6 show the ground truth.

Method	Pre. Dataset	Backbone	DDD17	DSEC
	Training from scratch			
ResNet [23]	-	ResNet50	56.96	57.60
	Transfer learning			
ESS [61]	-	-	61.37	53.29
ResNet [23]	ImageNet	ResNet50	59.25	58.50
	Pre-training on N-ImageNet			
ECDP [76]	N-ImageNet	ResNet50	59.15	59.16
MEM [33]	N-ImageNet	dVAE+VIT	-	44.62
DMM [28]	N-ImageNet	-	60.59	58.78
Ours	N-ImageNet	ResNet50	61.98	61.07
Ours (w/ HF)	N-ImageNet	ResNet50	62.13	61.29
	Pre-training on E-TartanAir			
ECDDP [77]	E-TartanAir	Swin_T	62.56	61.25
Ours	E-TartanAir	Swin_T	63.16	61.77
Ours (w/ HF)	E-TartanAir	Swin_T	**63.29**	**62.05**

Table 6. Quantitative comparison of semantic segmentation on DDD17 and DSEC datasets. We report the mean intersection over Union (mIoU, %) for each dataset.

4.4. Semantic Segmentation

Since the N-ImageNet [32] dataset contains limited motion and scene information, we follow the approach in [77], and add a new set of experiments by using a synthetic event dataset, E-TartanAir [77], based on the TartanAir [71] dataset for pretraining. Detailed information can be found in the **Supplement**. We fine-tune the pretrained STP with image pretrained models on a downstream semantic segmentation task, attaching the UperNet decoder [3, 74] to our pretrained model to estimate semantic labels. We conducted experiments on the DDD17 [1, 5] and DSEC datasets [21, 61], using mean Intersection over Union (mIoU) as the evaluation metric. We included prior SOTA models ESS [61] in the comparison. Additionally, we integrated hierarchical features from STP into the semantic segmentation pipeline via a linear layer (w/ HF). More details can be found in the **Supplement**.

As shown in Tab 6, our method pretrained on E-TartanAir achieves the best results on both datasets, surpassing the SOTA ESS. Incorporating hierarchical features from STP improved the model's ability to capture temporal information in the event stream, further enhancing segmentation performance. We present the semantic segmentation results in Fig. 6. For more details and training parameters, please refer to the **Supplement**.

Method	Backbone	indoor_flying1		indoor_flying2		indoor_flying2	
		AEE	Outlier	AEE	Outlier	AEE	Outlier
Previous SOTA method							
EST [20]	-	1.24	5.09	2.05	19.90	1.71	11.67
DCEFlow [68]	-	0.75	0.60	1.39	8.01	1.13	5.29
Transfer learning from models pretrained on ImageNet [13]							
ViT [15]	ViT-S/16	0.88	3.06	1.79	16.63	1.49	8.66
ResNet [23]	ResNet50	0.60	0.23	1.37	8.76	1.15	5.34
Pretraining on N-ImageNet [32] + Finetuning							
ECDP [76]	ResNet50	0.60	0.35	1.35	8.57	1.12	5.26
ECDP [76]	ViT-S/16	0.61	0.05	1.26	6.69	1.00	3.11
Ours	ViT-S/16	0.58	0.05	1.22	6.34	0.93	3.03
Pre-training on E-TartanAir + Finetuning							
ECDDP [77]	Swin_T	0.36	0.04	0.45	0.002	**0.42**	**0.001**
Ours	Swin_T	**0.31**	**0.03**	**0.41**	**0.001**	0.43	**0.001**

Table 7. Comparison of optical flow estimation on the MVSEC dataset [85]. The evaluation following the KITTI benchmark [45].

4.5. Optical Flow Estimation

Event cameras excel at capturing dynamic data, making motion information measurement a crucial downstream task. To ensure a fair comparison with previous work, we also evaluated a model pretrained on E-TartanAir using STP. We replaced the classification head in the pretrained network with a decoder network for optical flow estimation [3, 25] and assessed its performance on the MVSEC dataset [85]. Details can be found in the **Supplement**. As shown in Tab. 7, our model pretrained on E-TartanAir [77] achieves the lowest Average Endpoint Error (AEE) and outlier ratio. Additionally, we provide sample optical flow predictions in the **Supplement**, further demonstrating the strong generalization performance of our approach.

5. Conclusion

In this paper, we propose a Spatiotemporal Information Fusion Prompting (STP) method that bridges event stream data with image pretrained models, enabling efficient transfer of pretrained knowledge. STP progressively integrates the spatiotemporal information of event data through a multi-scale spatiotemporal receptive field, addressing the sparsity issue of event data and making it compatible with image pretrained models for efficient image-to-event knowledge transfer. Moreover, STP follows the lightweight design principle of prompt-tuning, reducing 90% of pretraining parameters and requiring only 1/3 of the training epochs to complete pretraining efficiently. Experimental results demonstrate the superiority and potential of our method, offering a novel perspective for event camera data pretraining.

Acknowledgments

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant (2025A1515011485), in part by the National Natural Science Foundation of China (62027804), and the Major Key Project of Pengcheng Laboratory (PCL2025A02).

References

[1] Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 8

[2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 3

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2, 8

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 2

[5] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd17: End-to-end davis driving dataset. *arXiv preprint arXiv:1711.01458*, 2017. 8

[6] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 1, 3

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[11] Wensheng Cheng, Hao Luo, Wen Yang, Lei Yu, and Wei Li. Structure-aware network for lane marker extraction with dynamic vision sensor. *arXiv preprint arXiv:2008.06204*, 2020. 6

[12] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023. 2

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2, 5, 8

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 4, 5, 7, 8

[16] Peiqi Duan, Zihao W Wang, Xinyu Zhou, Yi Ma, and Boxin Shi. Eventzoom: Learning to denoise and super resolve neuromorphic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12824–12833, 2021. 1

[17] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3867–3876, 2018. 1

[18] 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. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1, 3

[19] Yue Gao, Siqi Li, Yipeng Li, Yandong Guo, and Qionghai Dai. Superfast: 200x video frame interpolation via event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[20] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019. 3, 5, 7, 8

[21] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 8

[22] 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. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 4, 5, 7, 8

[24] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4918–4927, 2019. 2

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable

vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 2, 8

[26] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International conference on machine learning*, pages 2712–2721. PMLR, 2019. 2

[27] Yuhuang Hu, Tobi Delbruck, and Shih-Chii Liu. Learning to exploit multiple vision modalities by using grafted networks. In *European Conference on Computer Vision*, pages 85–101. Springer, 2020. 2

[28] 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. 2, 5, 6, 8

[29] Zhilin Huang, Quanmin Liang, Yijie Yu, Chujun Qin, Xiawu Zheng, Kai Huang, Zikun Zhou, and Wenming Yang. Bilateral event mining and complementary for event stream super-resolution, 2024. 1

[30] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3

[31] Jianhao Jiao, Huaiyang Huang, Liang Li, Zhijian He, Yilong Zhu, and Ming Liu. Comparing representations in tracking for event camera-based slam. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 1369–1376, 2021. 1

[32] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2146–2156, 2021. 1, 2, 5, 6, 8

[33] Simon Klenk, David Bonello, Lukas Koestler, Nikita Araslanov, and Daniel Cremers. Masked event modeling: Self-supervised pretraining for event cameras. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2378–2388, 2024. 2, 5, 6, 8

[34] Chankyu Lee, Adarsh Kumar Kosta, Alex Zihao Zhu, Kenneth Chaney, Kostas Daniilidis, and Kaushik Roy. Spike-flownet: event-based optical flow estimation with energy-efficient hybrid neural networks. In *European Conference on Computer Vision*, pages 366–382. Springer, 2020. 1

[35] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 2, 3

[36] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3

[37] Quanmin Liang, Zhilin Huang, Xiawu Zheng, Feidiao Yang, Jun Peng, Kai Huang, and Yonghong Tian. Efficient event stream super-resolution with recursive multi-branch fusion. 1

[38] Quanmin Liang, Xiawu Zheng, Kai Huang, Yan Zhang, Jie Chen, and Yonghong Tian. Event-diffusion: Event-based image reconstruction and restoration with diffusion models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3837–3846, 2023. 1

[39] Patrick Lichtensteiner, Christoph Posch, and T Delbruck. A 128x128 120db 15μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, (2):566–576, 2008. 1

[40] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 2, 3

[41] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 3

[42] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7

[43] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 7

[44] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5419–5427, 2018. 3, 7

[45] M. Menze, C. Heipke, and A. Geiger. Joint 3d estimation of vehicles and scene flow. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W5: 427–434, 2015. 8

[46] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters*, 7(2):3515–3522, 2022. 2

[47] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017. 1

[48] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126(12):1381–1393, 2018. 1

[49] Xing Nie, Bolin Ni, Jianlong Chang, Gaofeng Meng, Chunlei Huo, Shiming Xiang, and Qi Tian. Pro-tuning: Unified prompt tuning for vision tasks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3

[50] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:159859, 2015. 6

[51] Federico Paredes-Vallés and Guido CHE De Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021. 1

[52] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11557–11568, 2021. 2

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[54] Aniruddh Raghu, Jonathan Lorraine, Simon Kornblith, Matthew McDermott, and David K Duvenaud. Meta-learning to improve pre-training. *Advances in Neural Information Processing Systems*, 34:23231–23244, 2021. 2

[55] Vignesh Ramanathan, Rui Wang, and Dhruv Mahajan. Predet: Large-scale weakly supervised pre-training for detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2865–2875, 2021. 2

[56] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 1, 7

[57] Catherine D Schuman, Shruti R Kulkarni, Maryam Parsa, J Parker Mitchell, Bill Kay, et al. Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science*, 2(1):10–19, 2022. 1

[58] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1731–1740, 2018. 6

[59] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 1

[60] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18043–18052, 2023. 1

[61] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022. 2, 7, 8

[62] Yiwen Tang, Ray Zhang, Zoey Guo, Xianzheng Ma, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Point-peft: Parameter-efficient fine-tuning for 3d pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5171–5179, 2024. 3

[63] Gemma Taverni, Diederik Paul Moeys, Chenghan Li, Celso Cavaco, Vasyl Motsnyi, David San Segundo Bello, and Tobi Delbruck. Front and back illuminated dynamic and active pixel vision sensors comparison. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(5):677–681, 2018. 1

[64] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022. 1

[65] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[66] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios. *IEEE Robotics and Automation Letters*, 3(2):994–1001, 2018. 1

[67] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. Spot: Better frozen model adaptation through soft prompt transfer. *arXiv preprint arXiv:2110.07904*, 2021. 3

[68] Zhexiong Wan, Yuchao Dai, and Yuxin Mao. Learning dense and continuous optical flow from an event camera. *IEEE Transactions on Image Processing*, 31:7237–7251, 2022. 8

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

[70] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14733–14743, 2022. 7

[71] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. 2020. 8

[72] Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. P2p: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. *Advances in neural information processing systems*, 35:14388–14402, 2022. 3

[73] Ziyi Wu, Xudong Liu, and Igor Gilitschenski. Eventclip: Adapting clip for event-based object recognition. *arXiv preprint arXiv:2306.06354*, 2023. 2, 3, 5, 6

[74] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 8

[75] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 2

[76] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10699–10709, 2023. 2, 5, 6, 7, 8

[77] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data dense pre-training. In *European Conference on Computer Vision*, pages 292–310. Springer, 2025. 2, 8

[78] Zhiyang Yu, Yu Zhang, Deyuan Liu, Dongqing Zou, Xijun Chen, Yebin Liu, and Jimmy S Ren. Training weakly supervised video frame interpolation with events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14589–14598, 2021. 1

[79] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022. 2

[80] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10042–10051, 2021. 2

[81] Xu Zheng and Lin Wang. Eventdance: Unsupervised source-free cross-modal adaptation for event-based object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17448–17458, 2024. 2, 3, 5, 6

[82] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Eventbind: Learning a unified representation to bind them all for event-based open-world understanding. 3, 5, 6

[83] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Exact: Language-guided conceptual reasoning and uncertainty estimation for event-based action recognition and more. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18633–18643, 2024. 2, 3

[84] Alex Zihao Zhu and Liangzhe Yuan. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. In *Robotics: Science and Systems*, 2018. 1, 3, 7

[85] 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. *IEEE Robotics and Automation Letters*, 3(3): 2032–2039, 2018. 8

[86] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 1, 3, 7

[87] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Point-clip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2639–2650, 2023. 3

[88] Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, and Ying Fu. Learning to reconstruct high speed and high dynamic range videos from events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2024–2033, 2021. 3, 7