Event-based Video Reconstruction Using Transformer

Wenming Weng Yueyi Zhang* Zhiwei Xiong
University of Science and Technology of China

Abstract

Event cameras, which output events by detecting spatio-temporal brightness changes, bring a novel paradigm to image sensors with high dynamic range and low latency. Previous works have achieved impressive performances on event-based video reconstruction by introducing convolutional neural networks (CNNs). However, intrinsic locality of convolutional operations is not capable of modeling long-range dependency, which is crucial to many vision tasks. In this paper, we present a hybrid CNN-Transformer network for event-based video reconstruction (ET-Net), which merits the fine local information from CNN and global contexts from Transformer. In addition, we further propose a Token Pyramid Aggregation strategy to implement multi-scale token integration for relating internal and intersected semantic concepts in the token-space. Experimental results demonstrate that our proposed method achieves superior performance over state-of-the-art methods on multiple real-world event datasets. The code is available at https://github.com/WarranWeng/ET-Net.

1. Introduction

Event cameras, also known as neuromorphic cameras [45], are novel bio-inspired visual sensors, providing researchers with a radically different sensing paradigm. Rather than directly reporting frame-based representation at a fixed rate in conventional cameras, event cameras are specifically designed to detect and record spatio-temporal changes for each pixel. Compared with frame-based counterparts, event cameras possess several superior properties: high temporal resolution (about 1 μs), high dynamic range (140 dB) and low power consumption (5 mW) [24], which are suitable in scenarios that are challenging for conventional cameras, such as HDR scenes and high speed moving scenes. However, the event streams are not convenient for observation and post-processing due to their sparse, irregular and unstructured properties. To better utilize the advantages of event cameras, an intuitive way is to convert event streams to video composed of sequential intensity frames, which serves as a bridge that connects the off-the-shelf frame-based algorithms [12, 13, 11, 26, 42, 41, 44, 43] to event cameras.

Deep learning techniques, especially convolutional neural networks, have achieved great successes in the area of computer vision. Recently, several works performed event-based video reconstruction via deep learning methods and demonstrated impressive performance. Using supervised learning, Rebecq *et al*. [29, 28] first proposed the E2VID CNN-based model, achieving significant performance boost in terms of image quality and temporal consistency against hand-crafted methods [4, 21, 32]. Based on E2VID, Scheerlinck *et al*. [33] reduced inference time and model capacity using a light-weight network FireNet with only a minor drop in accuracy incurring. Further, Stoffregen *et al*. [35] presented that these supervised training methods showed a strong dependence on the synthetic data generated by event camera simulators, such as ESIM [27]. Consequently, for relaxing this data dependency, Federico *et al*. [22] approached, for the first time, the reconstruction problem from a self-supervised learning perspective via combining estimated optical flow and the event-based photometric constancy to train neural networks without ground-truth.

These CNN-based architectures [29, 28, 35, 22] show the preponderance in video reconstruction for event cameras. However, classic CNN-based models are not capable of modeling the long-range dependency due to the essential locality of convolution operations. Actually, capturing long-range dependency plays a crucial role in deep neural networks for both sequential data in NLP tasks and image data in vision tasks. Especially, CNN-based models are not effective to deal with structures that show large internal variation in terms of texture, shape and size. In order to tackle this limitation, some works have been proposed recently. Wang *et al*. [39] proposed a non-local operation, which can be plugged into multiple existing CNN models. Schlemper *et al*. [34] integrated additive attention gate modules into the skip-connections for global contexts. More recently, *Transformer* [36], designed for sequence-to-sequence prediction, has emerged as a popular architecture in both NLP and vi-

*Correspondence should be addressed to zhyuey@ustc.edu.cn

sion tasks [36, 38, 6, 10, 7]. Built upon the self-attention mechanisms solely instead of CNNs, Transformer shows an appealing potential in modeling global context information.

In this paper, we present the first attempt that explores the application of Transformer in the context of high speed video reconstruction for event cameras. Based on the novel perspective of sequence-to-sequence prediction, we propose ***E**vent **T**ransformer **Net**work* (ET-Net) to exploit the powerful potentials of Transformer for reconstructing video from pure events. Different from previous works [10, 48], our ET-Net adopts a hybrid CNN-Transformer architecture to leverage both detailed multi-resolution spatial information from CNN features and the global context encoded by Transformer. We verify that the combination of localized features and global contexts is able to further promote the reconstruction quality. Additionally, we propose a novel *Token Pyramid Aggregation* (TPA) module to implement multi-scale token integration, which is a core component of ET-Net. The proposed TPA represents the 2-D features using visual tokens and learns to directly relate semantic concepts in token-space instead of convolution operators, yielding a better reconstruction accuracy. Extensive experiments conducted on the existing frequently-used event camera datasets show that our proposed architecture ET-Net outperforms existing CNN-based methods, substantiating effectiveness of our transformer-based method.

We summarize our contributions as three-fold. (**1**) We propose ET-Net, a novel hybrid CNN-Transformer framework, to leverage both fine local information from CNN and global context from Transformer for approaching the event-based video reconstruction task. (**2**) We propose a Token Pyramid Aggregation module to perform multi-scale token integration for relating internal and intersected semantic concepts in the token-space. (**3**) We comprehensively demonstrate the effectiveness of our architectural design via extensive experiments, achieving a substantial performance boost over CNN-based methods.

2. Related Work

2.1. Event-based video reconstruction

Video reconstruction is a popular topic in the event-based vision literature. Photometric constancy, which means each event provides one equation relating intensity gradient and optical flow, serves as an early attempt to approach event-based video reconstruction problem. Kim *et al*. [14] showed the first study to design an Extended Kalman Filter to reconstruct a gradient image and presented the feasibility to predict 6-DOF camera motions in their future work [15]. Using the primal-dual algorithm, Bardow *et al*. [4] simultaneously optimized both optic flow and intensity estimation through a sliding spatio-temporal window. Another parallel route of research is built upon direct event integration without assuming scene structure or motion dynamics. Reinbacher *et al*. [21] introduced direct integration with periodic manifold regularization on the Surface of Active Events and optimized an energy function to reconstruct video from events. Scheerlinck *et al*. [32] proposed complementary and high-pass filtering to achieve computationally efficient, continuous-time video reconstruction.

Recently, deep learning methods, especially convolutional neural networks (CNNs), have shown the potentials to solve the event-based video reconstruction problem. Wang *et al*. [37] and Pini *et al*. [25] utilized generative adversarial networks (GANs) to reconstruct intensity with real grayscale frames. Rebecq *et al*. [29, 28] presented a novel CNN-based model that was trained in a supervised manner with a large-scale synthetic dataset for promoting reconstruction quality. Scheerlinck *et al*. [33] proposed a light-weight framework to achieve significant acceleration in terms of inference time. More recently, Stoffregen *et al*. [35] proposed a novel strategy of reducing the *simulation-to-reality gap* in between the synthetic dataset and the realistic dataset, bringing a considerable performance boost on multiple datasets. Instead of only using CNN networks, in this paper, we present a new method to reconstruct video from pure events by formulating a Transformer-based framework, enabling us to synthesise higher quality video.

2.2. Transformer

Transformer was first proposed by Vaswani *et al*. [36] for machine translation and have dominated in various natural language processing tasks [9, 46, 8, 16] as a de-facto architecture. Transformer consists of multiple self-attention layers for modeling long-range dependency and aggregating global contexts across the whole sequence, which is analogous to Non-Local Neural Networks [39] but without any recurrence or convolution operators.

With the significant success of Transformer in NLP field, many explorations for applying Transformer in computer vision tasks have been made recently. Carion *et al*. [6] reformulated the object detection as a direct set prediction problem from a sequence-to-sequence view and proposed a novel end-to-end detection Transformer (DETR) to generate bounding boxes. Chen *et al*. [7] utilized pre-trained technique to maximally excavate the capacities of transformer and attained state-of-the-art performance in multiple low-level image processing tasks. Dosovitskiy *et al*. [10] proposed Vision Transformer (ViT) to conduct image recognition, showing that when coupled with pre-training on sufficient data, Transformer possesses solid advantages against convolution neural networks. To the best of our knowledge, no prior works develop Transformer model for event-based video reconstruction.

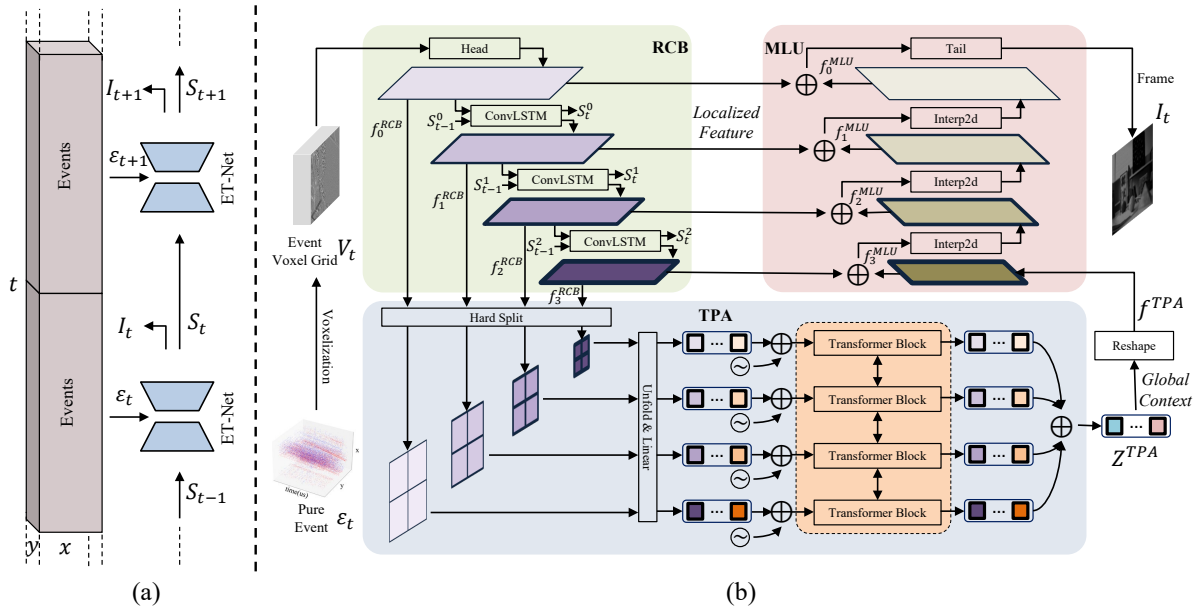


Figure 1. (a) The overview of our reconstruction framework. (b) The architecture of our proposed ET-Net. Generally, ET-Net is a U-shaped network, consisting of Recurrent Convolution Backbone (RCB), Token Pyramid Aggregation (TPA) and Multi-Level Upsampler (MLU). RCB extracts a feature pyramid from the event voxel grid. TPA further models internal and intersected long-range dependency from the feature pyramid and outputs the global context. Then MLU aggregates the localized feature from RCB and the global context from TPA to reconstruct the final intensity frame. The details of our network architecture including hyper-parameters and Transformer block design are elaborated in Sec. 3.2 and the supplementary material.

3. Proposed method

In this section, we present our transformer-based model ET-Net to approach this problem. Firstly, we introduce our strategy to generate fixed-size voxel grid for accommodating the processing fashion in canonical neural networks in Sec. 3.1. Subsequently, our proposed framework ET-Net and loss functions are elaborated in Sec. 3.2 and Sec. 3.3 respectively. The overall pipeline of our method is illustrated in Fig. 1.

3.1. Event Representation

The pure event stream $\mathcal{E} = \{e_{t_k}\}_{k=1}^{N_e}$, where N_e represents the number of events, is feed into our network. Each event $e_{t_k} \in \mathcal{E}$ is denoted as a four-element tuple (x_k, y_k, t_k, p_k), reporting spatial coordinates, timestamp and polarity respectively. In order to make the event stream compatible with the processing algorithms designed for frame-based vision, it is necessary to convert event stream $\mathcal{E}$ into a grid-like event voxel grid $V \in \mathbb{R}^{B\times H\times W}$ with B time bins via temporal bilinear interpolation [50]. Specifically, we perform this conversion according to

$$V(k) = \sum_i p_i \max(0, 1 - |k - \tfrac{t_i - t_0}{t_{N_e} - t_0}(B-1)|), \quad (1)$$

where t_0, t_{N_e} denote the start time and end time of event stream $\mathcal{E}$ respectively, $k \in [0, B-1]$. This converting method evenly populates the whole event stream to B consecutive and non-overlapping sections, of which each event contributes its polarity to two closest bins. In this work, we use $B = 5$ for conducting all experiments.

3.2. Event Transformer Network (ET-Net)

We propose a hybrid CNN-Transformer model ET-Net for event-based video reconstruction. Our model jointly exploits CNN and Transformer to produce localized features and global contexts respectively. The proposed ET-Net follows the classic encoder-decoder architecture. The input of our network is an event voxel grid V by populating a event stream $\mathcal{E}$ as described in Sec. 3.1. The output of our model is the final reconstructed intensity frame $\mathcal{I}$.

Recurrent Convolution Backbone (RCB). Instead of directly performing feature sequentialization on the event voxel grid V, we first feed it to a recurrent convolution backbone, which is composed of a head and three recurrent convolution blocks. The head is employed for transforming the input event voxel grid $V \in \mathbb{R}^{B\times H\times W}$ into the first scale feature $\mathbf{f}_0^{RCB} \in \mathbb{R}^{C_0\times H\times W}$. In our work, we set C_0 as 32.

Exploiting temporal consistency between successive

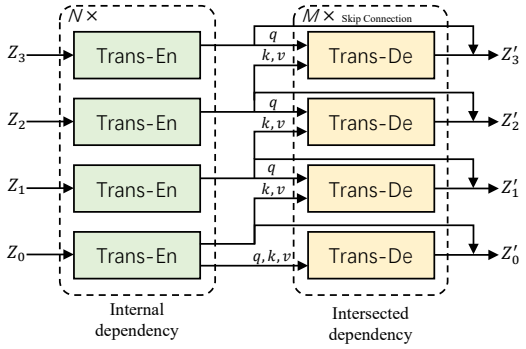


Figure 2. The detailed structure of Transformer Blocks utilized in TPA (four scales shown here). A single Transformer Block consists of N Transformer encoders and M Transformer decoders. Skip connection is employed to transfer the output tokens by Transformer encoder to the final tokens Z_l'.

frames benefits our video reconstruction for events. As in [29], a ConvLSTM layer is employed in each recurrent block, which utilizes the previous states to enhance the temporal stability of reconstruction. Furthermore, in each recurrent block, we apply a convolutional layer (stride to 2) to decrease the spatial size of features by half. Meanwhile, the channel number doubles with the increase of scale level, *i.e.* $C_l = C_0 \times 2^l$. Consequently, three stacked recurrent blocks produce feature maps at three scales, which can be formulated as

$$\mathbf{f}_l^{RCB}, \mathbf{s}_l^t = f_l^{rec}(\mathbf{f}_{l-1}^{RCB}, \mathbf{s}_l^{t-1}), \tag{2}$$

where $l \in \{1, 2, 3\}$ denotes the l^{th} layer, $\mathbf{s}_l^t$ denotes the state of l^{th} layer at time t. Through recurrent convolution backbone, we finally obtain a multi-scale feature pyramid $\{\mathbf{f}_l^{RCB} \mid l \in \{0, 1, 2, 3\}\}$, which is passed to token pyramid aggregation module and multi-level upsampler module subsequently. Please see Fig. 1 for more details of RCB.

Token Pyramid Aggregation (TPA). For Vision Transformer, prior works [10, 48] generate single scale tokens via performing image sequentialization on input images. Similarly, built upon the features by CNN backbone, DETR [6] models the long-range dependency on the last scale feature (with small spatial size) by ResNet50, which loses the intersected correlation and spatial details from the other scales. Actually multi-scale aggregation has shown the excellent promotion in many vision tasks [18, 31, 30]. Therefore, we design the Token Pyramid Aggregation module, which is based on Transformer, to model the internal and intersected dependency from the feature pyramid extracted by RCB.

First, sequentialization operation is performed on each feature in the feature pyramid extracted by RCB. Specifically, we divide the feature $\mathbf{f}_l^{RCB} \in \mathbb{R}^{C_l \times \frac{H}{2^l} \times \frac{W}{2^l}}$ into small patches. The dimension of each patch in the l^{th} scale is $C_l \times \frac{P}{2^l} \times \frac{P}{2^l}$ ($P = 8$ in our work). Thus for each scale, we have the same number ($\frac{HW}{P^2}$) of small patches. Then we flatten these patches into one-dimensional vectors, forming a sequence $\{\mathbf{f}_{l,i}^P \in \mathbb{R}^{\frac{P^2 C_l}{4^l}} \mid i \in \{0, .., \frac{HW}{P^2} - 1\}\}$. We further apply a linear projection f_l^{proj} and sinusoidal positional encoding [36] $\mathbf{e}_i \in \mathbb{R}^D$ to map each patch $\mathbf{f}_{l,i}^P$ into a latent one-dimensional embedding token $\mathbf{T}_{l,i} \in \mathbb{R}^D$, which is formulated as

$$\mathbf{T}_{l,i} = f_l^{proj}(\mathbf{f}_{l,i}^P) + \mathbf{e}_i. \tag{3}$$

Please refer to [36] for the reason of positional encoding if needed, which is not claimed here for brevity. The illustrative process of sequentialization operation can be found in the supplementary material. After patch embedding and positional encoding, we reformulate the token sequence $\{\mathbf{T}_{l,i}\}$ to a token matrix $Z_l \in \mathbb{R}^{\frac{HW}{P^2} \times D}$, which can be processed by Transformer Blocks subsequently. In our work, we set D as 256.

As shown in Fig. 2, for each scale, one Transformer block is employed to model both internal dependency and intersected dependency from the feature pyramid. Within each Transformer block, we stack several vanilla Transformer encoders, which extract the internal dependency of tokens in each scale via the self-attention operations. Then Transformer decoders are appended to build the intersected dependency on adjacent scale tokens. Note that the key and value vectors fed to Transformer decoders are from the Encoder of the Transformer Block in the lower scale, while the query vector are still from the encoder of current Transformer Block. This design endows our network with the capacity of learning to extract and exchange multi-scale contexts, which can be further demonstrated in Sec. 4.3.

Additionally, we also introduce the residual connection to maintain the internal dependency by Transformer encoders. Specifically, for each scale, the output tokens from Transformer encoder and Transformer decoder are added as Z_l' via skip connection. We then aggregate all tokens $\{Z_l'\}$ from different scales to generate the hidden token matrix Z^{TPA} for TPA. Note that Z^{TPA} shares the same dimension as Z_l. The details of Transformer Block that are utilized in our work are illustrated in Fig. 2.

Multi-Level Upsampler (MLU). TPA outputs a two-dimensional matrix $Z^{TPA} \in \mathbb{R}^{\frac{HW}{P^2} \times D}$, the resolution of which is not the same as the original resolution. Therefore, we design a Multi-Level Upsampler to recover the full resolution intensity image $\mathcal{I} \in \mathbb{R}^{H \times W}$ from the hidden tokens Z^{TPA} and feature pyramid $\{\mathbf{f}_l^{RCB} \mid l \in \{0, ..., 3\}\}$. Notably, the hidden tokens capture the long-range dependency across the multi-scale feature set, while the feature pyramid provides localized information. These two data streams merit both CNN and Transformer, significantly enhancing the reconstruction quality compared with using only one of them, which can be further demonstrated in Sec. 4.3.

Specifically, MLU consists of three stacked upsampling blocks and a tail (please see Fig. 1 for more details). Each upsampling block is built with a bi-linear interpolation operation followed by a convolutional layer, where the upsampling factor is 2 to maximally alleviate the adversarial effect. Before feeding the token matrix $Z^{TPA} \in \mathbb{R}^{\frac{HW}{P^2} \times D}$ to MLU, we reshape it to a three-dimensional feature $\mathbf{f}^{TPA} \in \mathbb{R}^{D \times \frac{H}{P} \times \frac{W}{P}}$. Formally, the calculation process of upsampling blocks can be formulated as

$$\mathbf{f}_l^{MLU} = \begin{cases} f_{up}(\mathbf{f}_{l+1}^{MLU} + \mathbf{f}_{l+1}^{RCB}), l = 0, 1, 2 \\ \mathbf{f}^{TPA}, l = 3 \end{cases} \tag{4}$$

where $\mathbf{f}_l^{MLU}$ denotes the feature generated by upsampling block in the l^{th} scale. The tail is a simple convolution layer, which takes the combination of $\mathbf{f}_0^{MLU}$ and $\mathbf{f}_0^{RCB}$ features as input (please see the skip connection between RCB and MLU in Fig 1), generating the final reconstructed intensity image $\mathcal{I} \in [0, 1]^{H \times W}$.

3.3. Loss functions

We employ LPIPS and temporal consistency loss functions for training, which are also adopted in [29]. The LPIPS loss is a differentiable similarity metric to evaluate the frame quality. The temporal consistency loss measures a photometric error between two aligned successive reconstructed images, which is used to mitigate the temporal artifacts.

The final loss $\mathcal{L}$ over T time-steps can be calculated as

$$\mathcal{L} = \sum_{t=0}^{T} \mathcal{L}_R^t + \lambda_{TC} \sum_{t=L_0}^{T} \mathcal{L}_{TC}^t, \tag{5}$$

where $\mathcal{L}_R^t$, $\mathcal{L}_{TC}^t$ are the LPIPS reconstruction loss and temporal consistency loss at time t, L_0 denotes the starting index for computing temporal consistency loss and λ_{TC} controls the temporal consistency proportion in the final loss. We set T, L_0 and λ_{TC} to 40, 2 and 1 respectively.

4. Experiments and results

4.1. Experimental setup

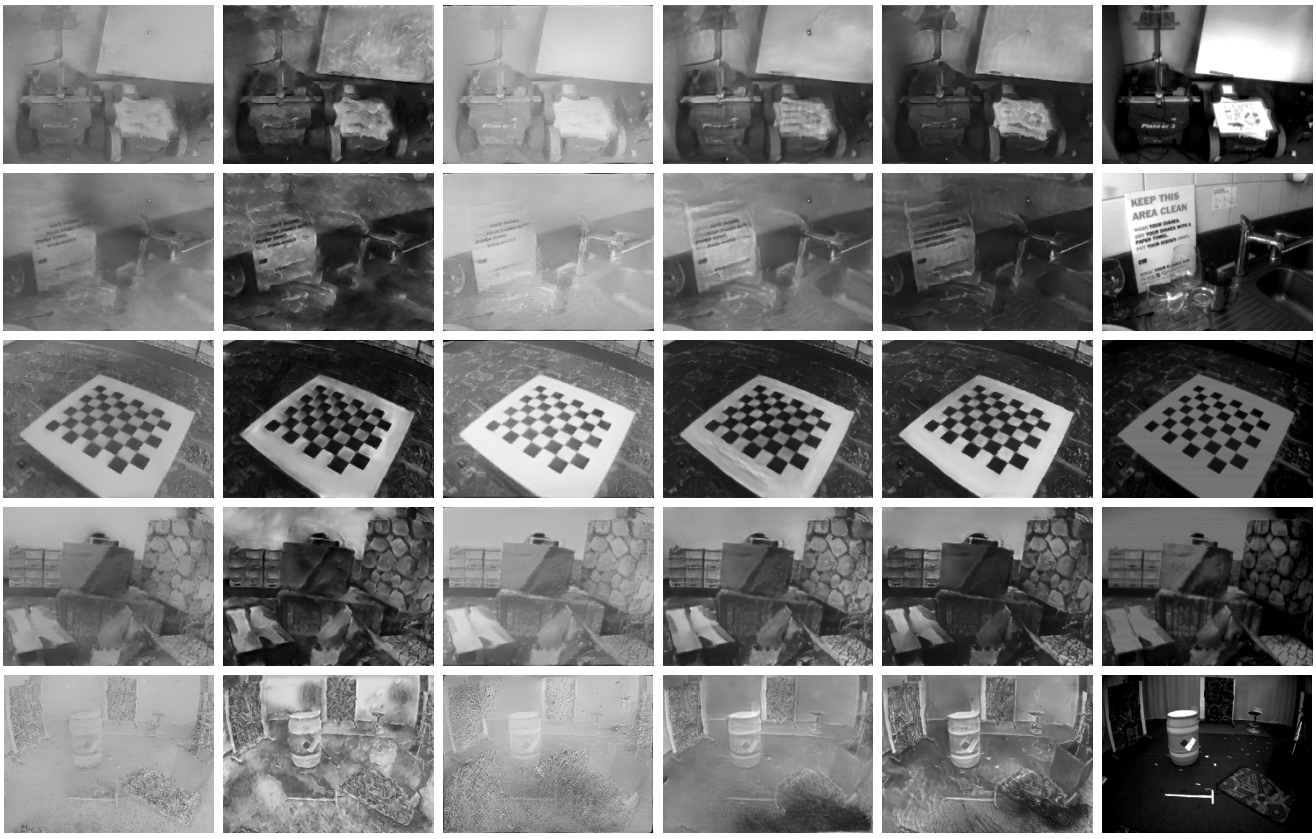
Training dataset. Our proposed network performs video reconstruction from pure events in a supervised manner. A large number of event sequences with corresponding ground-truth frames are indispensable for training. In order to make fair comparison, we follow the same generation scheme as E2VID+ [35] to synthesize the training dataset via ESIM[27], an excellent simulator for synthesizing events with reliable ground-truth frames. Specifically, "Multi-Object-2D" is adopted as the rendering mode to run the simulation, which enables the foreground multi-objects moving across a background image with various 2-D motion properties in terms of translations, rotations and dilation. We combine the object images provided by [35] with images from the COCO dataset [17] to make the candidate foreground multi-objects. The background images are randomly chosen from the COCO dataset too. We launch the simulation procedure by endowing each image with random trajectories. The contrast thresholds (CTs) are picked between 0.1 and 1.5 in an ascending order, of which positive CTs and negative CTs are restricted to the limitation: $\mathrm{C_p} = \mathrm{C_n} \times x, x \in \mathcal{N}(\mu = 1.0, \sigma = 0.1)$. The whole training dataset contains 280 sequences with 256×256 resolution. Each sequence lasts 10 seconds.

The training data augmentation strategy is the same as [35]. Specifically, Gaussian noise $\mathcal{N}(\mu = 0, \sigma = 0.1)$ is added to the input event tensor for simulating the background noise. A few 'hot' pixels that fire spurious events are also simulated. We perform random cropping with size of 112×112 and random flipping for the input event tensor. Additionally, we also employ random pause augmentation. Please refer to [35] for the pause augmentation.

Testing datasets. We evaluate our model on three publicly released event-based datasets: HQF [35], IJRR [20] and MVSEC [49]. The HQF dataset, recorded by two DAVIS240C [5] cameras, provides high quality ground-truth frames, of which the motion blur is maximally mitigated under preferable exposure. 14 sequences are contained , covering a wider range of motions and scene types, including static scenes and motion scenes of slow, medium and fast, indoor and outdoor scenes. IJRR provides 25 realistic datasets by DAVIS240C [5] and two synthetic datasets via the event camera simulator. MVSEC is recorded by a synchronized stereo event camera system. Each sequence of MVSEC releases extensive ground-truth reference data for evaluations. Compared with HQF, IJRR and MVSEC are not designed specifically for the event-based video reconstruction problem. For fair comparison, we select the same sequences from the two datasets as those reported in [35]. The exact cut times of IJRR and MVSEC sequences can be found in the supplementary material.

Evaluation metrics. For quantitative evaluation, we consider three widely-used evaluation metrics: (i) mean squared error (MSE), (ii) structure similarity (SSIM) [40] and (iii) perceptual similarity (LPIPS) [47], which are also utilized in [28, 35, 33]. A lower value of MSE and LPIPS or a higher value of SSIM indicates a better performance.

Implementation details. Our network is implemented using the Pytorch framework [23]. AdamW [19] is utilized as the optimizer with the initial learning rate 0.0002. We adopt an exponential decay strategy of learning rate with gamma of 0.99. Our model is trained for 300 epochs with batch size of 2 on 2 NVIDIA Tesla V100 GPUs.



FireNet FireNet+ E2VID E2VID+ ET-Net (Ours) Ground-truth

Figure 3. Qualitative comparison with baseline methods on HQF (Row 1&2), IJRR (Row 3&4) and MVSEC (Row 5). Our proposed network demonstrates better reconstruction results with fine-grained details and minor artifacts, while other baselines present foggy effects across the whole image, which induces a severe brightness disturbance. More visual results can be found in the supplementary material.

Methods	MSE ↓			SSIM ↑			LPIPS ↓		
	HQF	IJRR	MVSEC	HQF	IJRR	MVSEC	HQF	IJRR	MVSEC
FireNet	0.0981	0.1333	0.287	0.522	0.488	0.247	0.467	0.338	0.718
FireNet+	0.0465	0.0568	0.228	0.595	0.535	0.265	0.326	0.298	0.574
E2VID	0.1824	0.1830	0.313	0.477	0.448	0.227	0.515	0.357	0.727
E2VID+	0.0371	0.0650	0.135	0.638	0.551	0.337	**0.258**	0.241	0.513
Ours	**0.0349**	**0.0503**	**0.113**	**0.643**	**0.585**	**0.358**	0.274	**0.237**	**0.491**

Table 1. Quantitative comparison of baseline methods of event-based video reconstruction on HQF, IJRR and MVSEC. Best in bold, the second best with underline. The breakdown results can be found in the supplementary material.

4.2. Comparison with the state-of-the-art methods

We compare our proposed method with four state-of-the-art methods FireNet [33], FireNet+ [35], E2VID [28] and E2VID+ [35]. FireNet+ and E2VID+, which share the same architecture with FireNet and E2VID, are retrained using the synthetic training dataset [35].

For all state-of-the-art methods, we perform evaluations using the pre-trained model obtained from [1, 2, 3]. For fair comparison, we keep all experiment settings the same. No post-processing operations (such as grayscale normalization and histogram equalization) are performed for all the methods. Table 1 shows the quantitative comparison results. In terms of MSE, our ET-Net outperforms FireNet+ and E2VID+ by 15 % over all three datasets, which is a solid improvement. As for SSIM, our ET-Net surpasses E2VID+ with a clear margin, achieving 0.643, 0.585 and 0.358 on HQF, IJRR, and MVSEC respectively. In terms of LPIPS, ET-Net still matches or exceeds the state-of-the-art method E2VID+ except for a minor drop on HQF. The

Model	MSE ↓	SSIM ↑	LPIPS ↓
ET-Net-2-s4	**0.113**	**0.376**	**0.494**
E2VID-res6	0.165	0.319	0.536
ET-Net-4-s4	**0.118**	**0.355**	**0.491**
E2VID-res12	0.169	0.309	0.521
ET-Net-6-s4	**0.167**	**0.312**	0.538
E2VID-res16	0.180	0.311	**0.518**

Table 2. Ablation results of ET-Net and E2VID variants on MVSEC. Each pair of models has similar parameter amounts. The difference lies in whether Transformer-based TPA is utilized in the model.

quantitative comparison of each scene of three datasets are provided in the supplementary material. It should be noted that we achieve the best performance when $N = 3, M = 2$ in Transformer Blocks and three scales are aggregated in TPA. The total parameter amount of this ET-Net is 22M.

Figure 3 illustrates the qualitative results reconstructed by our ET-Net and all baseline methods on images of video clips from the HQF, IJRR and MVSEC datasets. The ground-truth images are also listed for comparison. It can be observed that FireNet and E2VID reconstruct frames with higher intensity values, presenting foggy artifacts across the whole image plane. The reconstruction results of E2VID+ and FireNet+ are better than those of FireNet and E2VID in visual effects, presenting a more realistic scene. Our ET-Net further brings more detailed contexts to the final reconstructions, additionally reducing common failure cases like stretch marks witnessed in the FireNet+. The image contrast of our reconstructed frame is very close to that of the ground-truth image. These qualitative results support the quantitative results in Table 1. In the supplementary material, we also provide several reconstructed video clips and apply our model to the High Speed and HDR scenes.

4.3. Network architecture analysis

In order to investigate the importance of components in our ET-Net, we perform the ablation analysis under various settings, including: 1) TPA exists or not; 2) the number of scales in TPA; 3) the depth of Transformer block in each scale; 4) Transformer decoders in TPA exist or not; 5) skip connection between RCB and MLU. All the models are trained on our synthetic training dataset for 200 epochs with batch size of 8, and evaluated on all three testing datasets. Notably, unless specialized in the main paper, the other ablation results can be found in the supplementary material with additional clarifications. All other experimental settings stay the same as Sec. 4.1.

Before presenting the detailed ablation results, we describe a nomenclature for ET-Net variants. The names

Model	MSE ↓	SSIM ↑	LPIPS ↓
ET-Net-4-s4	**0.0552**	**0.587**	**0.236**
ET-Net-5-s3	0.0584	0.564	0.242
ET-Net-8-s2	0.0636	0.547	0.260
ET-Net-16-s1	0.0991	0.509	0.284

Table 3. Ablation results of ET-Net variants which have different aggregation scales in TPA on IJRR.

Model	MSE ↓	SSIM ↑	LPIPS ↓
ET-Net-2-s4	0.0413	0.619	0.288
ET-Net-4-s4	**0.0403**	**0.635**	**0.277**
ET-Net-6-s4	0.0430	0.623	0.286

Table 4. Ablation results of ET-Net variants which have different encoder and decoder numbers in Transformer Blocks on HQF.

of ET-Net variants follow the pattern "ET-Net-(A)-s(B)", where A represents the total number of Transformer encoders and decoders at each scale in TPA and B represents the aggregation scales in TPA. Please refer to our supplementary material for details.

TPA. Our proposed TPA is utilized to exploit the global context of event tensors, which is also the major contribution of our work. We design ablation experiments to validate the effectiveness of TPA. The basic structure of our ET-Net is similar to that of E2VID. Thus we compare ET-Net with E2VID and evaluate the performance quantitatively. We first choose three ET-Net variants with different configurations: ET-Net-2-s4, ET-Net-4-s4 and ET-Net-6-s4, which have different parameter amounts. For fair comparison, we revise the structure of E2VID via adding more residual blocks so that these E2VID variants have similar parameter amounts to their counterparts. Specifically, we construct three E2VID variants: E2VID-res6, E2VID-res12 and E2VID-res16, of which the Resblock number utilized in the model is 6, 12 and 16 respectively. Table 2 shows the quantitative results of three pairs of models on MVSEC. It can be seen that with our TPA module, the performance is improved to a large extent, which validates the effectiveness of TPA.

Aggregation scales in TPA. We further investigate the influences of TPA with different scales for ET-Net. For keeping the similar parameter number, we adopt ET-Net-4-s4, ET-Net-5-s3, ET-Net-8-s2 and ET-Net-16-s1 to conduct this ablation on IJRR. Notably, these ET-Net variants consist of similar number of Transformer encoders and decoders, with the stacking fashion as the main distinction. Table 3 presents that our ET-Net with multiple-scale TPA performs better than single scale variant (ET-Net-16-s1), although ET-Net-16-s1 variant possesses deeper layers for modeling global context.

Setting	IJRR			MVSEC			HQF		
	MSE ↓	SSIM ↑	LPIPS ↓	MSE ↓	SSIM ↑	LPIPS ↓	MSE ↓	SSIM ↑	LPIPS ↓
w/ Trans-Decoders	**0.0522**	**0.587**	**0.236**	**0.118**	**0.355**	**0.491**	0.0403	**0.635**	**0.277**
w/o Trans-Decoders	0.0563	0.572	0.245	0.129	0.341	0.498	**0.0399**	0.634	0.280

Table 5. Ablation results of the ET-Net-4-s4 model with and without Transformer decoders in TPA.

Setting	MSE ↓		SSIM ↑		LPIPS ↓	
	ET-Net-4-s4	E2VID-res12	ET-Net-4-s4	E2VID-res12	ET-Net-4-s4	E2VID-res12
w/ Skip Connection	**0.118**	**0.169**	**0.355**	**0.309**	**0.491**	**0.521**
w/o Skip Connection	0.132	0.179	0.320	0.273	0.653	0.716

Table 6. Ablation results of the ET-Net-4-s4 and E2VID-res12 models with and without skip connection on MVSEC.

Depth of Transformer block. We provide our investigation on the number of Trans-block in each scale for exploring appropriate number of encoders and decoders in each TPA scale. We choose ET-Net variants ET-Net-2-s4, ET-Net-4-s4 and ET-Net-6-s4 to perform ablation experiments on HQF. Table 4 shows the ablations results. It can be observed that small or large number of Transformer encoders and decoders cannot result in a satisfying performance. Our models with small capacity are not capable of capturing the long range dependency from the latent CNN features, while large models show overfitting and degrade the generalization ability. Therefore, we speculate that the best performance should be achieved near the place where the total number of Transformer encoders and decoders in Transformer blocks is 4. The ET-Net model on which we report the best performance in Sec. 4.2 has 5 encoders and decoders, which is consistent with previous speculation.

Transformer decoder in TPA. The TPA in our ET-Net possesses two kinds of Transformer components: Transformer encoder for modeling the internal dependency of tokens in each scale and Transformer decoder for building the intersected dependency across tokens from the adjacent scales. In order to investigate the influence of Transformer decoder in TPA, we replace the Transformer decoders with Transformer encoders of the same number to form a new ET-Net variant for conducting this ablation. We report the results on IJRR, MVSEC and HQF for the extensive comparison in Table 5. It can be observed that the utilization of Transformer decoders improves the reconstruction performance on all three datasets.

Skip connection. Our ET-Net leverages both the low-level precise details from CNN and the global contexts from Transformer. We merge the localized features into global tokens generated by Transformer progressively in MLU via skip connection. In order to determine the effect of localized features in ET-Net, we perform an ablation experiment via removing the skip connection and report the results in terms of MSE, SSIM and LPIPS. Moreover, we also perform this ablation on the E2VID model, which takes the skip connection following the UNet [31] design. Both the ET-Net-4-s4 and E2VID-res12 variants are employed to conduct the experiments, which share similar amount of parameters. This experiment is conducted on the MVSEC dataset. As shown in Table 6, the two variants with skip connection outperform those without skip connection, bringing an average improvement of 15%. It is worth noting that our ET-Net achieves better performances over E2VID both under w/ skip and w/o skip settings, which intensively reinforces the importance of global context in our ET-Net.

5. Conclusion

In this paper, we propose ET-Net, a novel Transformer-based framework, to approach the event-based video reconstruction problem for the first time. Coupling CNN with Transformer, ET-Net possesses the potentials to maximally excavate the respective advantages of CNN and Transformer. Additionally, we further propose the TPA module to perform multi-scale token integration. With the input from pyramidal low-level features extracted by CNN, TPA represents the 2-D features using visual tokens and learns to directly relate semantic concepts in token-space instead of convolution operators. Extensive experiments demonstrate that our proposed network achieves superior performances over state-of-the-art methods on multiple datasets, opening up a new avenue for event-based video reconstruction.

However, while ET-Net shows significant improvements, inference time and model memory consumption are occupied more than CNN-based models due to complex self-attention calculations in Transformers. In the future, we will further apply knowledge distillation and model pruning techniques to promote our model.

Acknowledgements. We acknowledge funding from National Key R&D Program of China under Grant 2017YFA0700800 and National Natural Science Foundation of China under Grant 61901435.

References

[1] https://www.cedricscheerlinck.com/firenet. 6

[2] http://rpg.ifi.uzh.ch/E2VID. 6

[3] https://timostoff.github.io/20ecnn. 6

[4] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 884–892, 2016. 1, 2

[5] 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. 5

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2, 4

[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 2

[8] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. 2

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 4

[11] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2019. 1

[12] 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

[13] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018. 1

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

[15] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *European Conference on Computer Vision*, pages 349–364. Springer, 2016. 2

[16] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 2

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 4

[19] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[20] 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. 5

[21] 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, 2

[22] 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 Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021. 1

[23] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[24] Lichtsteiner Patrick, Christoph Posch, and Tobi Delbruck. A 128x 128 120 db 15μ s latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-state Circuits*, 43:566–576, 2008. 1

[25] Stefano Pini., Guido Borghi., and Roberto Vezzani. Learn to see by events: Color frame synthesis from event and rgb cameras. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP*, pages 37–47. INSTICC, SciTePress, 2020. 2

[26] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 1

[27] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on Robot Learning*, pages 969–982. PMLR, 2018. 1, 5

[28] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3857–3866, 2019. 1, 2, 5, 6

[29] 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*, 2019. 1, 2, 4, 5

[30] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3577–3586, 2017. 4

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015. 4, 8

[32] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Asian Conference on Computer Vision*, pages 308–324. Springer, 2018. 1, 2

[33] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 156–163, 2020. 1, 2, 5, 6

[34] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53:197–207, 2019. 1

[35] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *European Conference on Computer Vision*. Springer, 2020. 1, 2, 5, 6

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 1, 2, 4

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

[38] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy, July 2019. Association for Computational Linguistics. 2

[39] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 1, 2

[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[41] Zeyu Xiao, Xueyang Fu, Jie Huang, Zhen Cheng, and Zhiwei Xiong. Space-time distillation for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2113–2122, 2021. 1

[42] Zeyu Xiao, Zhiwei Xiong, Xueyang Fu, Dong Liu, and Zheng-Jun Zha. Space-time video super-resolution using temporal profiles. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 664–672, 2020. 1

[43] Ruikang Xu, Zeyu Xiao, Jie Huang, Yueyi Zhang, and Zhiwei Xiong. Edpn: Enhanced deep pyramid network for blurry image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 414–423, 2021. 1

[44] Ruikang Xu, Zeyu Xiao, Mingde Yao, Yueyi Zhang, and Zhiwei Xiong. Stereo video super-resolution via exploiting view-temporal correlations. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2021. 1

[45] Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A dynamic vision sensor with 1% temporal contrast sensitivity and in-pixel asynchronous delta modulator for event encoding. *IEEE Journal of Solid-State Circuits*, 50(9):2149–2160, 2015. 1

[46] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5

[48] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021. 2, 4

[49] 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. 5

[50] 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 Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 3