IEEE Open Journal of Signal Processing

Received 4 February 2026; revised 30 April 2026; accepted 5 May 2026. Date of publication 14 May 2026; date of current version 28 May 2026. The review of this article was arranged by Associate Editor S. Wan.

Digital Object Identifier 10.1109/OJSP.2026.3693230

Deep Spatio-Temporal and Frequency Guided Fusion Network for Event-to-Video Reconstruction

RAMNA MAQSOOD [1], PAULO NUNES [1] (Member, IEEE), CAROLINE CONTI [1] (Member, IEEE), AND LUÍS DUCLA SOARES [1] (Senior Member, IEEE)

[1]Instituto de Telecomunicações, Instituto Universitário de Lisboa (ISCTE-IUL), 1649-026 Lisbon, Portugal

CORRESPONDING AUTHOR: RAMNA MAQSOOD (email: ramna.maqsood@lx.it.pt).

This work was supported in part by the National funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., and in part by EU funds through Project/support UID/50008/2025 –Instituto de Telecomunicações, with DOI identifier 10.54499/UID/50008/2025.

ABSTRACT Event-to-video (E2V) reconstruction has gained significant attention recently for its advantages in enabling high dynamic range and fast motion capture capabilities. However, event data encodes only relative brightness changes, lacking the absolute intensity information necessary for accurate reconstruction. Recent methods incorporate previously reconstructed images to provide intensity references but process them in the spatial domain where low- and high-frequency components are highly coupled. This spatial processing typically leads to the degradation of fine details and introduces artifacts such as over-smoothing, blurring and low contrast reconstruction. To address this, we propose a deep spatio-temporal and frequency guided fusion network for E2V reconstruction (DSTFN-E2V), featuring a dual-path architecture with two key components: i) a prior frequency decomposition module (PFDM), and ii) a spatio-temporal event-driven feature extraction module (STEM). The PFDM decouples low- and high-frequency information from previously reconstructed images and current event voxel grid via a 2D discrete wavelet transform, processing the low-frequency subband through residual blocks to preserve structural coherence and intensity references, while an edge-detail refinement module (ERM) enhances edge and texture details from high-frequency subbands. The frequency-specific features from PFDM and the spatio-temporal features from STEM are then integrated through the proposed event-image fusion blocks (EIFBs) that apply cross-attention across three encoder stages, enabling simultaneous structural preservation and detail recovery. Experiments on four real-world datasets demonstrate that DSTFN-E2V achieves state-of-the-art results with 12% SSIM improvements while being 50% faster than recent attention-based methods, with superior edge fidelity and reduced artifacts.

INDEX TERMS Event-to-video (E2V) reconstruction, discrete wavelet transforms, cross-attention networks.

I. INTRODUCTION

Event cameras are bio-inspired vision sensors that offer high temporal resolution and low latency, making them well-suited for scenes with fast motion. However, their asynchronous and sparse output is incompatible with most conventional computer vision algorithms designed for frame-based imagery. Event-to-video (E2V) reconstruction [1], [2], [3] addresses this gap by converting event streams into dense, frame-based representations, enabling the effective use of established computer vision models.

In recent years, deep learning (DL) has significantly advanced E2V reconstruction. Early works [4], [5], [6] introduced convolutional recurrent neural networks (CRNNs) that process event voxel grids to preserve temporal coherence. More recent works [7], [8], [9], [10] have begun utilizing previous reconstructions to guide current frame generation, recognizing that prior frames contain valuable information for enhancing reconstruction quality. Despite the significant improvements in reconstruction performance achieved by these DL strategies, existing methods face two fundamental

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

limitations that constrain reconstruction quality. Firstly, these methods perform reconstruction operations on events and previously reconstructed images solely in the spatial domain. In the spatial domain, high-frequency (HF) and low-frequency (LF) components are inherently coupled within pixel intensities, as each pixel value is simultaneously influenced by both LF components (such as smooth regions and global structure) and HF components (such as edges and fine details). This mixing makes it difficult to separately control how different frequencies are processed [11]. As a result, convolutional and self-attention (SA) operations tend to favor smooth, LF features over fine details [12], leading to blurred reconstructions with a lack of structural details. Secondly, existing methods [1], [13], [14] require a heavy architecture for global feature learning. To capture complex spatial relationships and global context, current approaches rely on computationally expensive attention-based models. This design choice becomes problematic when processing high-resolution event representations at inference time.

On the other hand, frequency-aware analysis methods, such as those based on the discrete wavelet transform (DWT), have shown strong potential in event-based vision. Unlike Fourier-based methods [2], [11] that decompose a signal into global frequency components without retaining information about where in the image those frequencies occur, DWT performs a spatially localized analysis by decomposing a signal into one LF and three HF subbands where each wavelet coefficient encodes both frequency content and its spatial location within the image. This property preserves the spatial correspondence between frequency components and image regions across multiple scales, which is particularly important in E2V reconstruction where precise spatial alignment between event-driven features and prior-frame frequency components must be maintained during fusion. The LF subband serves as a coarse approximation of the original image and preserves global structures and background context information that is typically absent in event data. On the other hand, the HF subbands capture sparse local edge details along different directions. By relying on these properties of DWT, some recent works have shown notable performance improvements in event-based vision [15], [16] by decomposing event representations into frequency domain at a lower computational cost [17]. Furthermore, operating on wavelet-decomposed subbands reduces the spatial resolution of feature maps by a factor of two (in each direction) at each decomposition level, directly reducing the computational cost of subsequent operations without sacrificing representational capacity.

In this work, we argue that decomposing a previously reconstructed image into frequency subbands and fusing them with event-driven features can offer two complementary benefits that are grounded in the physical properties of event cameras. First, since events are triggered by logarithmic intensity changes at pixel locations, event data is inherently sensitive to edges and motion boundaries which is precisely the information captured by HF wavelet subbands. Static edges that do not trigger events are preserved in the HF subbands, whereas motion-degraded edges that generate events can be sharpened using event guidance. Second, event data captures intensity changes associated with structural motion but lacks dense background semantics, such as, global structure, and overall intensity, while the LF subband representing a coarse-scale version of the reconstructed image, preserves these global appearance properties. This physical alignment between event data characteristics and wavelet subband properties establishes a natural and principled division of roles: event features emphasize dynamic structural cues, while the LF subband provides stable semantic context.

These observations inspired us to explore the integration of frequency domain information into the E2V reconstruction to improve the quality of reconstructed video frames. To address the limitations of spatial-domain coupling and heavy global modeling in existing methods, we propose a deep spatio-temporal and frequency guided fusion network (DSTFN-E2V), featuring a dual-path architecture. Unlike prior approaches that process previously reconstructed images as spatial signals, DSTFN-E2V first decomposes the previous reconstruction and current event voxel grid into frequency-specific components using 2D-DWT. This decomposition enables our model to disentangle global appearance information from fine structural details before fusion with event-driven features. Specifically, a prior frequency decomposition module (PFDM) processes LF subbands from both the prior reconstruction and the event voxel grid, preserving global structures and intensity information that is inherently missing from event streams. Meanwhile the HF subbands are refined using a dedicated edge-detail refinement module (ERM) to enhance edges and textures degraded by spatial-domain smoothing. In parallel, a spatio-temporal event-driven module (STEM) extracts spatial and temporally aggregated features from event representations. This complementary information is then progressively fused through event-image fusion blocks (EIFBs), which perform lightweight cross-attention between event-driven features and frequency-aware priors at multiple encoder-decoder stages. By decoupling frequency information prior to fusion, DSTFN-E2V enables balanced reconstruction of LF realism and HF details without relying on heavy attention-based global modeling. As a result, the reconstructed frames exhibit improved structural consistency and contrast, while maintaining high computational efficiency. To the best of our knowledge, this work is the first to explicitly reformulate prior-image guidance in E2V reconstruction as a frequency-driven problem, where frequency components of previously reconstructed images are decoupled, processed according to their distinct roles, and systematically fused with event-driven features. Our contributions are as follows:

- We propose a novel spatio-temporal and frequency-based framework that reformulates prior-image guidance as a frequency-driven problem. By decoupling frequency components of previously reconstructed images before fusion with event-driven features, this framework

effectively alleviates common reconstruction artifacts such as over-smoothing, blurring, and loss of fine details.
- We design frequency-aware and event-guided fusion modules, including PFDM, ERM and EIFB, which model the complementary roles of frequency-domain priors and event-driven features. These modules enable selective preservation of global appearance information and enhancement of structural details across multiple reconstruction stages.
- We achieve improved reconstruction quality with higher computational efficiency, outperforming recent attention-based E2V methods. Despite employing cross-attention for multi-stage feature fusion, the proposed method remains lightweight and suitable for high-resolution E2V reconstruction.

II. RELATED WORK

Event-based Video Reconstruction: Initial E2V reconstruction approaches relied on hand-crafted methods. A complementary filter technique [18] combined events with frames using high-pass and low-pass filtering, while [19] formulated the problem as solving a Poisson equation where events represent gradients of the log-intensity field. In recent years, DL has shown remarkable efficacy for E2V reconstruction. Rebecq et al. [4] were the pioneers in introducing a CRNN model called E2VID and Scheerlinck et al. [6] followed with a computationally lighter variant, "FireNet". Stoffregen et al. [5] further enhanced results through E2VID+ and FireNet+ by utilizing statistics from synthetic datasets to bridge the domain gap with real-world data. Weng et al. [13] introduced the first transformer-based model utilizing SA to capture global context across event sequences, while Ma et al. [14] integrated spatial and channel attention with deformable convolutions for richer feature representation. In contrast to methods that process events alone, recent approaches combine events with intensity images, recognizing that they provide complementary information, with intensity capturing static scene content and events excelling in dynamic regions. SPADE-E2VID [10] employed spatially adaptive denormalization layers, to provide information from previous reconstructions into the network. HyperE2VID [9] pioneered the use of hypernetworks to dynamically modulate convolutional weights based on previous reconstructions. CISTA-LSTC [8] instead formulated reconstruction as an unrolled optimization problem, maintaining temporal consistency through learned sparse coding with previous reconstructions as initialization points. Another line of works [7], [12] first reconstructs an initial image via pre-trained models, then uses it as a conditional input alongside events, particularly within diffusion-based frameworks.

Deep Frequency-enhanced Models: Recent advances in DWT-enhanced DL models have attracted considerable attention for event-based vision. Fang et al. [20] proposed 3DWSNet for action recognition, decomposing event streams into frequency subbands to capture multi-scale temporal patterns. Cui et al. [15] introduced WD-DETR for object detection, employing wavelet denoising to separate signal from noise in event representations. Cheng et al. [21] developed a wavelet-based framework extracting frequency-decomposed features for recognition tasks. Sun et al. [16] explored cross-modal fusion for motion deblurring, applying DWT to both blurred images and events to align HF edge information. Leng et al. [17] theoretically established the computational advantage of wavelet transforms over learnable frequency representations, demonstrating that fixed wavelet basis functions eliminate trainable parameters while providing natural multi-resolution decomposition. Moreover, our prior works [22], [23] have shown the effectiveness of frequency-domain analysis in E2V reconstruction. While existing methods apply frequency analysis directly to event representations, they mainly operate on the event stream itself. In contrast, we apply frequency decomposition to both previously reconstructed image and current event voxel grid and introduce a dual-path fusion architecture that processes frequency subbands through specialized modules before integrating them with event features.

III. PROPOSED METHODOLOGY

This paper considers the problem of reconstructing a sequence of intensity images, $\hat{I}_k$, from an event stream. Starting from the common approach in the literature [9], [24] the event stream is converted into voxel grids. Each voxel grid, V_k, is then passed to the proposed network, DSTFN-E2V, which reconstructs a 2D grayscale image from it, having dimensions $H \times W \times 1$, where H and W represent the image height and width, respectively.

A. EVENT REPRESENTATION

In this section, we describe our methodology for representing event data as a voxel grid. Given an event stream, $\{e_i\}$, containing N_E events, each event, $e_i = (x_i, y_i, t_i, p_i)$, corresponds to the i-th event characterized by spatial coordinates, (x_i, y_i), timestamp, $t_i \in [0, T]$, and polarity, $p_i \in \{+1, -1\}$, indicating an increase $(+1)$ or decrease (-1) in brightness, for all $i \in \{0, \ldots, N_E\}$. We construct event voxel grids using a between-frames strategy in which events are grouped such that all events occurring between two consecutive frames are aggregated into the same voxel grid. Specifically, given the image timestamps, the set of events in the k-th group is defined as $E_k = \{e_i \mid T_{k-1} < t_i \leq T_k\}$ where T_k is the ending timestamp of the k-th group, and its duration is $\Delta T = T_k - T_{k-1}$. We divide the events in each group into B temporal bins to discretize the continuous-time event data; and adopted a common value of $B = 5$. The timestamps of the events in the group are first normalized to the range $[0, B-1]$. Each event then contributes its polarity to the two closest bins using bilinear interpolation. The voxel grid, $V_k \in \mathbb{R}^{H \times W \times B}$, for the k-th group is formed using (1):

$$V_k(x, y, t) = \sum_i p_i \max(0, 1 - |t - t_i^*|)\delta(x - x_i, y - y_i) \tag{1}$$

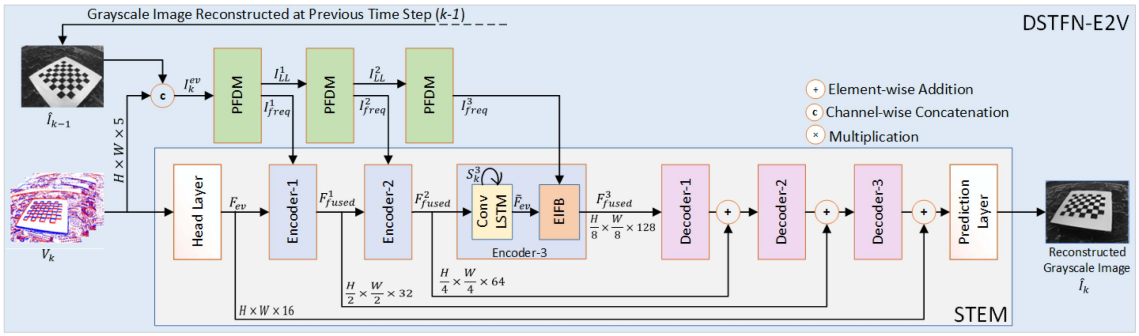


FIGURE 1. Overall architecture of the proposed DSTFN-E2V.

where δ is the Kronecker delta that is used to select the event location and $t_i^* = (B-1)\frac{t_i - T_k}{\Delta T}$ is the normalized timestamp.

B. PROPOSED DSTFN-E2V ARCHITECTURE

The proposed DSTFN-E2V utilizes a dual-path architecture designed to process the event voxel grid, V_k, at each time step, k, producing a corresponding reconstructed image, $\hat{I}_k$, (as shown in Fig. 1). The main path, known as the STEM, consists of three encoder-decoder stages with skip connections. In each encoder, convolutional log-short term modules (ConvLSTMs) are used to capture the spatio-temporal dependencies in V_k. In parallel, the second path, PFDM, processes both the previous reconstructed image, $\hat{I}_{k-1}$, and the current voxel grid V_k. PFDM extracts frequency-aware features that describe both the static and dynamic elements of the scene. The spatio-temporal features from STEM and frequency-aware features from PFDM are then fused at each encoder stage through the proposed EIFBs, which apply cross-attention to integrate global context from both paths into the final reconstructed image, $\hat{I}_k$.

Recurrent Formulation: Let S_k^l denote the recurrent state of the l-th ConvLSTM [4] encoder block in the STEM path at time step k (as shown in Fig. 1). Given the states from the previous time step, $\{S_{k-1}^l\}_{l=1}^3$, and the current event voxel grid, V_k, our network predicts the intensity image, $\hat{I}_k$, and computes the current states, $\{S_k^l\}_{l=1}^3$, as (2):

$$(\hat{I}_k, \{S_k^l\}_{l=1}^3) = \text{STEM}(V_k, \{S_{k-1}^l\}_{l=1}^3, \{I_{\text{freq}}^l\}_{l=1}^3) \quad (2)$$

where I_{freq}^l denotes the prior frequency features at decomposition level l, which are generated by the PFDM and fed into the corresponding encoder stage of STEM as (3):

$$I_{\text{freq}}^l = \text{PFDM}(V_k, \hat{I}_{k-1}), \quad l \in \{1, 2, 3\} \quad (3)$$

1) PRIOR FREQUENCY DECOMPOSITION MODULE (PFDM):

The input to the PFDM is the concatenation of current event voxel grid, $V_k \in \mathbb{R}^{H\times W\times 5}$, and the previously reconstructed image, $\hat{I}_{k-1} \in \mathbb{R}^{H\times W\times 1}$, forming $I_k^{\text{ev}} \in \mathbb{R}^{H\times W\times 6}$.

2D Discrete Wavelet Transform (2D-DWT): We apply a three-level hierarchical 2D-DWT decomposition to progressively separate frequency components at multiple scales. At each level, the 2D-DWT decomposes the input into four

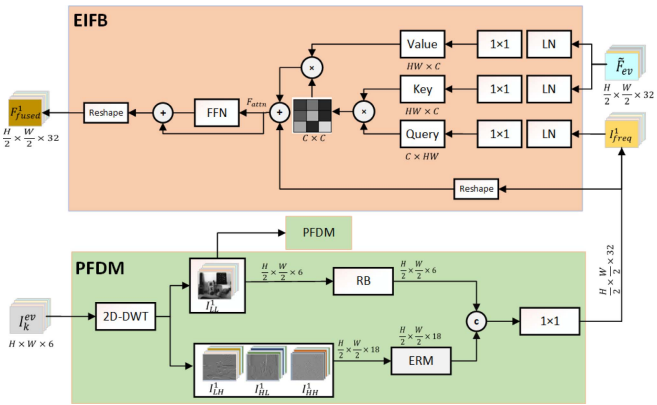


FIGURE 2. Detailed illustration of the PFDM and EIFB. The illustrated module structures and spatial dimensions correspond to the first level of decomposition; processing at other scales follows the same design.

subbands at half spatial resolution: one LF approximation subband, $I_{LL}^l \in \mathbb{R}^{(H/2^l)\times(W/2^l)\times 6}$, and three HF detail subbands, $I_{LH}^l, I_{HL}^l, I_{HH}^l \in \mathbb{R}^{(H/2^l)\times(W/2^l)\times 6}$. This hierarchical decomposition applies 2D-DWT recursively on the LL subband from the previous level, producing features at resolutions $H/2^l \times W/2^l$ (Fig. 2). The LL subband captures coarse-scale structural information through low-pass filtering in both horizontal and vertical directions, encoding global scene layout and intensity distribution. The LH subband isolates horizontal edge information (high-pass horizontal, low-pass vertical), HL captures vertical edges (low-pass horizontal, high-pass vertical), and HH encodes diagonal edge details (high-pass in both directions). This decomposition separates semantic context from edge-centric features in a single operation [17]. We use the Haar wavelet [25] for this decomposition. We experimented with other wavelet bases, including Daubechies-2 and symlets, but observed no significant difference in reconstruction performance. Given this, we selected the Haar wavelet due to its lower computational complexity arising from its simple filter structure. We evaluated different decomposition depths (2, 3, and 4-levels). Moving from 2 to 3 levels yields consistent improvement across all metrics, while further increasing to 4-levels provides only marginal gains with additional computational cost. Therefore, we adopt

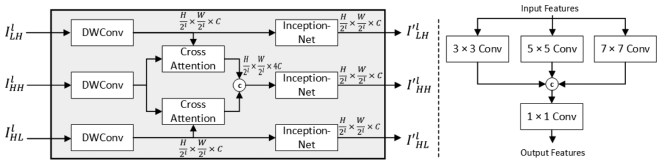


FIGURE 3. Detailed illustration of the ERM module (left) and the layered architecture of the Inception-Net (right).

a 3-level decomposition as a trade-off between performance and efficiency. We found that this separation enables three complementary mechanisms for cross-modal fusion: i) HF subbands explicitly encode edge and intensity transition features that align with the event-driven intensity changes, enabling structural correspondence between frequency-domain features and event representations; ii) HF subbands preserve both static and dynamic edge structure, complementing events that only capture moving edges; and iii) LL subband provides absolute intensity and semantic context that events lack. While DWT provides frequency separation, raw wavelet coefficients require adaptive processing to extract task-relevant features. CNNs have showed superior capability in learning hierarchical feature representations from wavelet subbands [26], [27] as convolutional filters can adaptively emphasize informative patterns while suppressing noise within each frequency band. Therefore, at each decomposition level, the LL subband, I^l_{LL}, passes through a residual block (RB) consisting of two 3×3 convolutional layers (Conv) with ReLU activation and skip connection to preserve structural information.

Edge-detail Refinement Module (ERM): The three HF subbands at each level, l, contain directional edge information critical for reconstruction. However, directly using raw HF coefficients can introduce noise and inconsistent edge responses due to sparsity in V_k and reconstruction errors in $\hat{I}_{k-1}$. Therefore, inspired by [28], we designed the ERM, where, as shown in Fig. 3, three HF subbands are initially processed through depthwise separable convolutions (DWConv) to extract features from each subband independently and efficiently. Then, we employ two cross-attention layers [29], where I^l_{HH} serves as query (Q), while I^l_{LH} and I^l_{HL} provide keys (K) and values (V), respectively. Cross-attention is preferred over CNN-based fusion as the contribution of horizontal and vertical edge features to diagonal edge refinement varies spatially depending on local content. Unlike fixed convolutional kernels, cross-attention allows I^l_{HH} to dynamically attend to the most relevant features from I^l_{LH} and I^l_{HL} at each spatial location, enabling adaptive directional complementarity. This mechanism also facilitates noise suppression by reducing the influence of inconsistent or spurious edge responses while preserving coherent edge structures. Finally, each enhanced subband then passes through a multi-scale parallel convolution block inspired by the inception architecture [30]. This block consists of three parallel branches with different kernel sizes convolutions, which simultaneously capture edge features at different spatial scales. The larger kernels provide broader contextual information, while smaller kernels preserve fine edge details. The outputs from all three branches are concatenated and fused through a convolution to produce the aggregated features. The processed LF and HF features in PFDM are concatenated and passed through a 1×1 Conv layer to produce the final frequency-decomposed features, I^l_{freq}.

2) SPATIO-TEMPORAL EVENT-DRIVEN MODULE (STEM):

Head Layer: The input $V_k \in \mathbb{R}^{H \times W \times 5}$ passes through a Conv layer with 3×3 kernel and stride 1 to extract initial features, F_{ev}, while preserving spatial dimensions.

Encoder Blocks: STEM employs three encoder stages, each containing a ConvLSTM followed by an EIFB. ConvLSTM first applies a Conv layer with 3×3 kernel and stride 2 to extract local features and downsample the spatial resolution. This downsampling matches the spatial dimensions of I^l_{freq} for cross-modal fusion. The convolved features then pass through an LSTM [31] to capture long-range temporal dependencies by maintaining hidden states across time steps. The output of this block, $\tilde{F}_{ev}$ is then fused with I^l_{freq} features through EIFB producing multi-scale fused representations, F^l_{fused}.

Event-Image Fusion Block (EIFB): While attention mechanisms have shown strong performance in E2V reconstruction [13], [14], standard SA incurs substantial computational cost due to pairwise computation of attention scores across all spatial locations, resulting in $O(H^2W^2)$ complexity. Furthermore, existing methods [8], [10] process event and image features through concatenation or process it only in spatial domain, limiting the network's ability to efficiently learn cross-modal dependencies. To address this, following the approach [32], our EIFB performs efficient cross-modal attention where frequency-aware features from the PFDM serve as Q to retrieve relevant information from event-driven features. Specifically, as shown in Fig. 2, both input features undergo layer normalization (LN) and 1×1 convolution to project them into a common embedding space of dimension C. The frequency features, I^l_{freq}, generate Q, while event features, $\tilde{F}_{ev}$, generate K and V. To reduce computational complexity, we perform attention in the channel dimension rather than the spatial dimension. We reshape Q, K, V from $\mathbb{R}^{C \times H \times W}$ to $\mathbb{R}^{C \times HW}$ and compute the channel-wise cross-attention as (4):

$$F_{\text{attn}} = V \cdot \text{softmax}\left(\frac{Q^T K}{\sqrt{d_k}}\right) \tag{4}$$

where $Q^T K \in \mathbb{R}^{C \times C}$. This design produces an attention map of size $C \times C$ instead of $HW \times HW$. Since C is a design parameter chosen independently of the input spatial resolution, the proposed formulation allows control over the model complexity. As a result, the computational cost no longer scales quadratically with the input spatial resolution, reducing from $O(H^2W^2)$ to $O(C^2)$. The attention mechanism performs soft indexing of K by Q features in the channel space, allowing HF components to selectively emphasize event channels encoding

edge information, while LF components attend to channels carrying broader temporal patterns. The frequency features, I^l_{freq}, are reshaped and combined with attention output, $F_{\text{attn}} \in \mathbb{R}^{C\times HW}$, via a residual connection. The result then passes through a feed-forward network (FFN) consisting of two linear layers with GELU activation applied after the first layer. The final fused features combine frequency-specific intensity information from previous reconstructions with temporal dynamics from events, enabling the decoder to utilize both static scene structure and dynamic changes for high-quality reconstruction.

Decoder Blocks: To progressively reconstruct the intensity image, we employ bilinear upsampling with the a factor of 2. These interpolated features are then refined through a Conv layer with 3×3 kernel size, allowing the decoder to capture local spatial relationships and further enhance feature representation.

Prediction Layer: The prediction layer consists of a 1×1 Conv layer that maps the features to the final intensity image. This yields the reconstructed image, $\hat{I}_k \in \mathbb{R}^{H\times W\times 1}$. We do not use an activation function after this layer.

C. TRAINING DETAILS

To train our model, we employ a combination of two loss functions. Firstly, we use learned perceptual image patch similarity (LPIPS) loss as employed in [4]. LPIPS measures the difference between the reconstructed image, $\hat{I}_k$, and the corresponding GT image, I_k, by comparing deep feature representations extracted using a VGG network [33] pre-trained on ImageNet [34]. Formally, LPIPS loss is defined as $\mathcal{L}^{LP}_k = j(\hat{I}_k, I_k)$, where j denotes the LPIPS distance. Secondly, we incorporate a temporal consistency loss, $\mathcal{L}^{TC}_k$, proposed in [4], to enforce smooth transitions between consecutive images in the reconstructed image sequence given as: (5):

$$\mathcal{L}^{TC}_k = M^k_{k-1} \left\| \hat{I}_k - \mathcal{W}^k_{k-1}(\hat{I}_{k-1}) \right\|_1 \tag{5}$$

where $M^k_{k-1} = \exp(-\alpha \|I_k - W^k_{k-1}(I_{k-1})\|^2_2)$ is a weighting term that accounts for occlusions, and $W^k_{k-1}(\hat{I}_{k-1})$ is the result of warping the reconstructed image at time step $k-1$ to time step k, using the optical flow. In our experiments, we set $\alpha = 50$ as done in [4], [5], [9]. The final loss function, $\mathcal{L}$, is computed using (6):

$$\mathcal{L} = \sum_{k=1}^{L} \mathcal{L}^{LP}_k + \lambda_{TC} \sum_{k=L_0}^{L} \mathcal{L}^{TC}_k \tag{6}$$

In our experiments, we set $L = 20$, which denotes the sequence length of consecutive images considered by our model during training. This length controls how many time steps the model unrolls to capture temporal dependencies. We set $\lambda_{TC} = 5$ and $L_0 = 2$, as indicated in [4].

We implement the proposed model in PyTorch for training. The network is trained over 300 epochs with a batch size of 4 on an RTX 3080 GPU with 16 GB of RAM. We utilize the Adam optimizer with an initial learning rate of 10^{-4}, which decays by 10% every 50 epochs. We also perform data augmentation with random crop and flip operations. The size of the crop was set to 128×128, and the probability of both vertical and horizontal flips was set to 0.5. To further enhance performance and mitigate overfitting, we incorporate noise augmentation, pause augmentation, and hot-pixel augmentation following the methods in [5].

During early training, the network's previous reconstruction, $\hat{I}_{k-1}$, used as input to PFDM, contains significant errors, which can destabilize learning. To address this, we adopt a curriculum learning (CL) scheme [35] from [9], where the network gradually transitions from GT supervision to self-generated priors. Specifically, we initially provide PFDM with GT previous images, I_{k-1}, rather than the actual network prediction, $\hat{I}_{k-1}$. Over the first 100 epochs, we linearly increase the contribution of predicted frames through a weighting schedule as (7) and (8):

$$\beta = \min\left(1, \frac{\text{epoch}}{100}\right) \tag{7}$$

$$I^{ev}_k = \beta \cdot \hat{I}_{k-1} + (1-\beta) \cdot I_{k-1} \tag{8}$$

where I^{ev}_k is the actual input to PFDM. This schedule allows the network to learn stable frequency decomposition on reliable intensity references before adapting to its own predictions. After epoch 100, we use only network predictions ($\beta = 1$), ensuring the model learns to refine its own outputs during inference when GT priors are unavailable.

IV. EXPERIMENTAL RESULTS

Training Dataset: We follow a common approach to generate a synthetic training dataset [5] using the "Multiple-Object-2D" rendering engine of ESIM [36]. This engine simulates multiple foreground objects moving across a background image, employing various 2D motion properties. Background images were selected from the MSCOCO dataset [37], while foreground objects were sourced from [5]. The dataset consists of 280 sequences, each 10 seconds in length. The contrast threshold values for event generation ranged from 0.1 to 1.5. Each sequence in our training dataset comprises an event stream together with the corresponding GT images and optical flow maps produced at an average rate of 51 Hz. The resolutions of event and image cameras are both 240×180 pixels.

Testing Datasets: To evaluate our proposed model, we utilize sequences from four real-world datasets: i) The event camera dataset (ECD) [38], recorded with a DAVIS240 C sensor at a resolution of 240×180; ii) The high-quality frames (HQF) dataset [5], recorded using two DAVIS240C sensors at the resolution of 240×180 pixels; iii) The multi vehicle stereo event camera (MVSEC) dataset [39] captured using a DAVIS346B sensor at a resolution of 346×260 pixels; and iv) A high-speed and HDR dataset [4], recorded with a Samsung DVS Gen3 camera with a resolution of 640×480 pixels, which includes challenging scenarios with strong light

TABLE 1. Quantitative Reconstruction Quality and Computational Efficiency Results

Type	Methods	ECD			HQF			MVSEC			ECD		Time(ms)	
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	RMS Contrast ↑	Params (M)	240 × 180	640 × 480
CNN-based	E2VID	12.64	0.459	0.321	10.54	0.461	0.372	6.93	0.241	0.642	36.941	10.712	1.92	7.52
	E2VID+	13.45	0.485	0.315	12.83	0.417	0.315	9.18	0.176	0.599	35.452	10.712	1.92	7.52
	FireNet	12.33	0.477	0.327	13.99	0.396	0.444	6.93	0.205	0.708	32.611	**0.042**	**0.98**	**2.76**
Prior-Guided	SPADE-E2VID	11.94	0.442	0.393	11.22	0.403	0.508	8.79	0.213	0.599	41.108	12.493	3.89	26.28
	HyperE2VID	13.11	0.513	0.301	12.61	0.495	0.356	9.62	0.278	0.573	40.652	10.152	2.19	13.35
	CISTA-LSTC	13.04	0.536	0.324	13.99	0.514	0.333	6.32	0.226	0.618	49.962	1.405	1.99	3.18
Attn-Based	ET-NET	14.52	0.548	0.296	14.29	0.502	0.312	10.15	0.311	0.569	47.082	22.182	10.35	112.32
	RecNet	11.57	0.491	0.305	12.58	0.426	0.371	9.65	0.226	0.621	38.681	11.853	9.85	98.63
Attn + Prior	DSTFN-E2V	**16.96**	**0.613**	**0.224**	**15.99**	**0.589**	**0.261**	**12.45**	**0.398**	**0.487**	**54.692**	9.112	3.15	39.90

contrasts and is used to demonstrate the generalizability of our model on a higher-resolution sensor.

Quality Evaluation Metrics: To quantitatively assess the proposed model, the following commonly used full reference metrics were used: peak signal-to-noise ratio (PSNR) ↑, structural similarity index measure (SSIM) ↑ and LPIPS ↓. Additionally, we employ root mean square (RMS) ↑ contrast, also used by prior works [2], [10]. The RMS contrast metric is used to evaluate image contrast by calculating the standard deviation of pixel intensities.

Benchmark Methods: We compare DSTFN-E2V against state-of-the-art (SOTA) methods across three categories: i) CNN-based (E2VID [4], E2VID+ [5] FireNet [6]) that process events through spatial convolutions and temporal aggregation without prior-guide; ii) prior-guided methods (SPADE-E2VID [10], HyperE2VID [9] CISTA-LSTC [8]) that incorporate $\hat{I}_{k-1}$ through spatial-domain modulation; and iii) attention-based methods (ET-Net [13], RecNet [14]). To ensure a fair evaluation, only methods with publicly available source code were selected for comparison. All methods were retrained on our dataset using identical settings and publicly available code, except for [14], which we implemented based on the model description provided in the original paper and have named as "RecNet". As the GT images of ECD are darker, we follow a common approach [4], [8] of normalizing the images into the range of [0, 1] to enable comparison with reconstructed images. As the GT images of ECD are darker, we follow a common approach [4], [8] of normalizing the images into the range of [0, 1] to enable comparison with reconstructed images.

Reconstruction Quality Results: Table 1 provides the average values of each metric across all evaluated frames. On ECD, our method surpassed the best SOTA method (ET-Net) by 2.44 dB PSNR and 0.065 SSIM, while reducing LPIPS to 0.224. Similar improvements are observed on HQF (15.99 dB PSNR, +1.70 dB over ET-Net) and MVSEC (12.45dB PSNR, +2.30 dB over ET-Net). Notably, DSTFN-E2V also achieves superior results over prior-guided methods that operate solely in the spatial domain, underscoring the effectiveness of frequency-aware feature modeling. Moreover, the improved RMS contrast computed for ECD, reflects DSTFN-E2V's ability to preserve global intensity distribution. Our LL subband processing maintains absolute intensity reference from previous reconstructions, preventing the intensity compression observed in methods without prior guidance (e.g., E2VID, FireNet: RMS contrast < 37).

Computational Efficiency Results: Table 1 also reports the computational efficiency in terms of the number of parameters (in millions) and average reconstruction time per frame (in milliseconds) at two different resolutions. Compared to existing attention-based methods, DSTFN-E2V is significantly lightweight, requiring fewer parameters while achieving lower inference times. This efficiency mainly stems from the design of the PFDM, which utilizes fixed wavelet-based components integrated with lightweight convolutional processing (see Section V). In addition, the proposed EIFB operates on compact feature interactions rather than dense spatial attention, further lowering inference overhead (see suppl. document). As a result, DSTFN-E2V achieves a favorable balance between reconstruction quality and efficiency, making it a well-suited attention-based method for resource-constrained event-based vision applications.

Visual Quality Results: To support quantitative results, we provided visual comparison in Fig. 4. As shown in the ECD results DSTFN-E2V better preserves fine structural details and sharp edges, which is consistent with its higher SSIM scores. In contrast, competing approaches exhibit noticeable blurring and loss of edge definition, particularly around object boundaries (see red rectangles in the first row). On the more challenging MVSEC dataset (third row), existing methods suffer from degraded textures under motion and lighting variations. DSTFN-E2V produces visually sharper reconstructions with clearer object structures (highlighted in red rectangle). For HQF dataset (last two rows), DSTFN-E2V successfully recovers fine textures and small structural details that are largely lost in competing methods (blue rectangles).

Fast Motion Scenarios Results: Fig. 5 shows DSTFN-E2V robustness under fast camera motion from the ECD dataset. While rapid movements typically cause severe motion blur and ghosting artifacts in E2V reconstruction, our method maintains scene detail recovery with minimal blurring.

High-resolution Reconstruction Results: To evaluate the generalization ability of our method at higher resolutions, Fig. 6 demonstrates that DSTFN-E2V produces competitive reconstructions comparable to existing methods while maintaining computational efficiency.

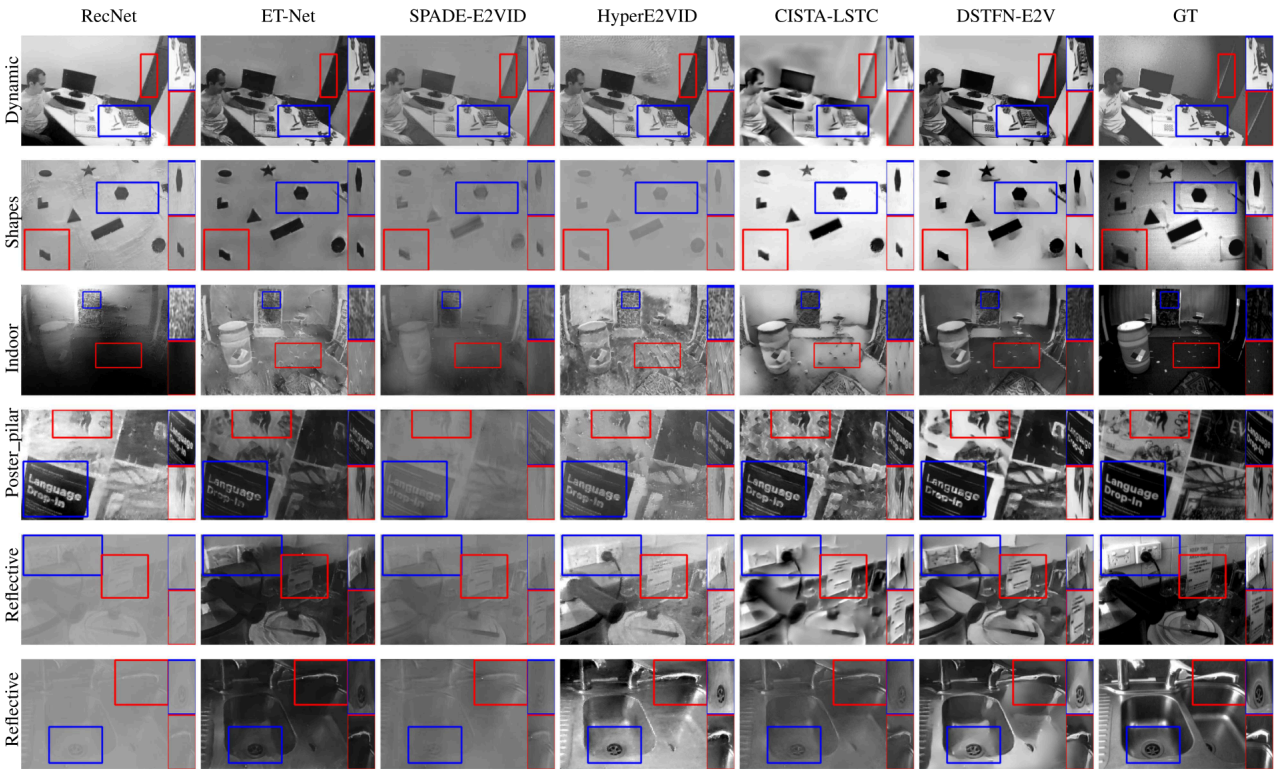


FIGURE 4. Visual quality results on three real-world event datasets. First two rows correspond to ECD, third for MVSEC and last three rows are from HQF dataset. Overall, our proposed DSTFN-E2V excels in detail recovery with minimum artifacts.



FIGURE 5. Visual quality results of DSTFN-E2V for dynamic_6dof of the ECD FAST subset.

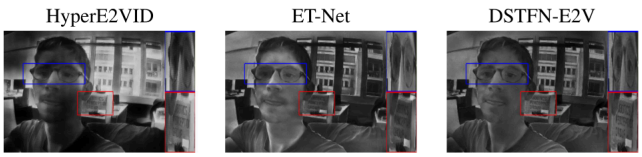


FIGURE 6. Visual quality results for hdr_selfie of the high-resolution HDR dataset.

V. ABLATION STUDY

In this section, we evaluate various design elements of the DSTFN-E2V model to verify their impact on performance.

w/o PFDM: We train a variant without the PFDM path, processing only V_k, through STEM. Table 2 shows, compared with the full model, removing PFDM results in significant decrease in performance (from both with and without CL). Notably, PFDM achieves better results only with 893 thousand additional parameters, providing an efficient balance between quality and computational cost. Fig. 7 shows top k-channels visualization from Encoder-2 of STEM. Without PFDM, channels appear relatively sparse, with most of the activations showing high contrast or very localized structures, whereas with PFDM, the top k-channels are much more consistent in their structure. The activations are smoother and less noisy, indicating that the model is focusing on more coherent patterns.

STEM w/ $\hat{I}_{k-1}$: We train a model that concatenates $\hat{I}_{k-1}$ directly with V_k in STEM to distinguish the benefits of our architectural features from the simple use of past reconstructions. This variant shows only marginal improvement compared to the model without PFDM, indicating that simple concatenation fails to integrate features from both modalities effectively.

PFDM w/o V_k: In this variant, only $\hat{I}_{k-1}$ is used for frequency decomposition, without incorporating the current V_k. The consistent decrease in metrics, especially SSIM and LPIPS, highlights that without V_k, the model fails to fully capture both static and dynamic features of the scene. This

TABLE 2. Ablation Results in Terms of Quality Evaluation Metrics and Number of Model Parameters Across Three Datasets. Results are Reported for Models Trained With and Without Curriculum Learning (CL) Strategy.

Variant	CL	ECD			HQF			MVSEC			Params (M)
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
i) w/o PFDM	✗	14.67	0.556	0.275	15.21	0.551	0.291	11.89	0.371	0.521	**8.21**
ii) STEM w/ $\hat{I}_{k-1}$	✗	14.70	0.558	0.279	15.24	0.553	0.294	11.92	0.374	0.524	8.21
iii) PFDM w/o V_k	✗	14.98	0.567	0.262	15.48	0.563	0.278	12.11	0.381	0.508	9.10
iv) w/o EIFB	✗	14.90	0.568	0.264	15.41	0.561	0.281	12.05	0.379	0.511	8.35
v) w/o ERM	✗	14.94	0.572	0.259	15.45	0.564	0.277	12.09	0.383	0.506	8.23
vi) PFDM w/o DWT	✗	14.72	0.559	0.273	15.26	0.554	0.289	11.94	0.373	0.519	9.11
DSTFN-E2V	✗	15.14	0.589	0.251	15.71	0.578	0.268	12.28	0.391	0.498	9.11
DSTFN-E2V	✓	**16.96**	**0.613**	**0.224**	**15.99**	**0.589**	**0.261**	**12.45**	**0.398**	**0.487**	9.11

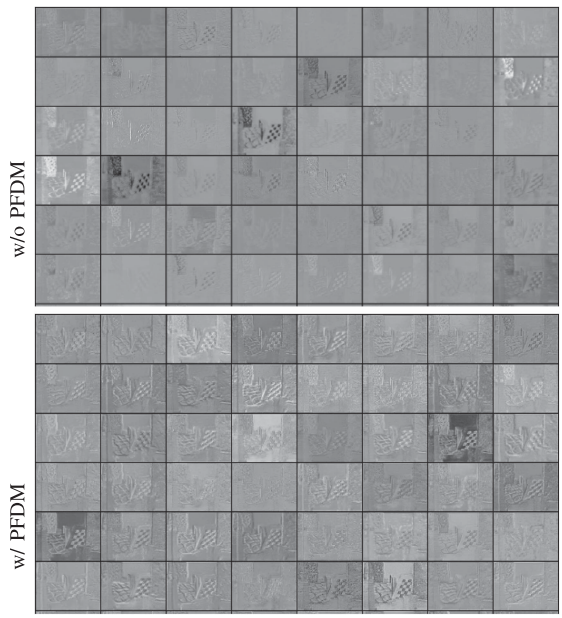


FIGURE 7. Top 48 channels from Encoder-2.

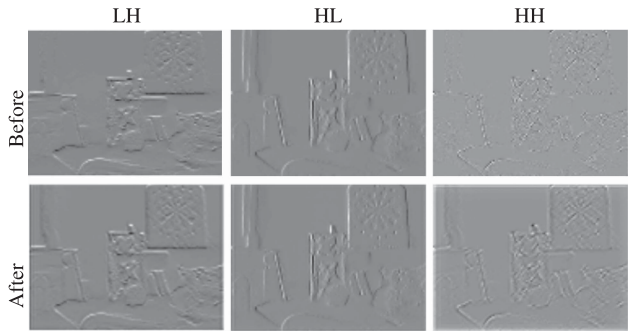


FIGURE 8. Visual effect of ERM on HF subbands.

lack of current event information limits the model's ability to effectively learn frequency-aware features from both the dynamic changes and the static components.

w/o EIFB: This variant replace EIFBs with simple concatenation followed by Conv layers at three encoder stages. This variant shows a notable degradation compared to the full model which shows that CNNs, which capture only local spatial relationships, cannot effectively model the global dependencies between I^l_{freq} and $\tilde{F}_{ev}$. In contrast, our EIFB employs cross-attention where I^l_{freq} features serve as Q and $\tilde{F}_{ev}$ features as K/V, enabling adaptive, long-range feature interactions. This mechanism allows the model to selectively attend to relevant event information based on frequency content, resulting in superior reconstruction quality.

w/o ERM: We train a model, where ERM is replaced with standard RBs. The 0.20 dB PSNR and 0.017 SSIM drop shows that ERM's cross-attention mechanism between directional subbands is crucial for preserving edge and texture details. Fig. 8 illustrates the effect of ERM processing: before ERM, the HF subbands contain noisy, fragmented edge information with inconsistent directional features; after ERM, the subbands exhibit enhanced edge clarity.

PFDM w/o DWT: To isolate the contribution of the wavelet-based frequency decomposition within PFDM, we train a variant in which the 2D-DWT is removed. In this variant, the concatenated input features are passed directly to RB and ERM without frequency decomposition. As shown in Table 2, this variant yields a notable performance drop compared to the full model, with results approaching those of the w/o PFDM variant. This confirms that without explicit frequency decoupling, the HF and LF components remain entangled within pixel intensities, preventing the model from separately controlling structural coherence and edge enhancement.

VI. CONCLUSION

This paper presented DSTFN-E2V, a novel deep spatial-temporal and frequency guided fusion network that addresses fundamental limitations in E2V reconstruction by decoupling the processing of frequency information. Our approach utilizes frequency decomposition through 2D DWT to enable independent handling of structural and fine detail recovery. The proposed PFDM provides global structure through LL subband processing, while the ERM enhances edge and texture details from HF subbands. These frequency-aware features are integrated with event-driven features through cross-attention based EIFB across multiple scales, effectively mitigating common artifacts such as over-smoothing,

blurring, and contrast degradation. Moreover, DSTFN-E2V provides almost 50% computational efficiency when compared with existing attention-based methods. Future directions include extending DSTFN-E2V with learnable frequency decomposition methods such as trainable wavelet transforms to further enhance reconstruction quality.

REFERENCES

[1] G. Guo et al., "Recurrent event-guided multimodal fusion for high dynamic range video reconstruction," *Expert Syst. Appl*, vol. 296, 2026, Art. no. 128988.

[2] J. Yu, X. Lu, L. Guo, C. Wang, G. Li, and J. Qian, "Event-based video reconstruction via spatial–temporal heterogeneous spiking neural network," *IEEE Trans. Circuits Syst. Video Technol*, vol. 35, no. 9, pp. 8478–8494, Sep. 2025.

[3] P. Gao, L. Wang, S. Ao, Y. Zhang, and Y. Guo, "Enhancing event-based video reconstruction with bidirectional temporal information," *IEEE Trans Multimedia*, vol. 27, pp. 4831–4843, 2025.

[4] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 43, no. 6, pp. 1964–1980, Jun. 2021.

[5] T. Stoffregen et al., "Reducing the sim-to-real gap for event cameras," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 534–549.

[6] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. E. Mahony, and D. Scaramuzza, "Fast image reconstruction with an event camera," in *Proc. Winter Conf. Appl. Comput. Vis.*, 2020, pp. 156–163.

[7] Q. Liang, X. Zheng, K. Huang, Y. Zhang, J. Chen, and Y. Tian, "Event-diffusion: Event-based image reconstruction and restoration with diffusion models," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 3837–3846.

[8] S. Liu and P. L. Dragotti, "Sensing diversity and sparsity models for event generation and video reconstruction from events," *IEEE Trans Pattern Anal. Mach. Intell*, vol. 45, no. 10, pp. 12444–12458, Oct. 2023.

[9] B. Ercan, O. Eker, C. Saglam, A. Erdem, and E. Erdem, "HyperE2VID: Improving event-based video reconstruction via hypernetworks," *IEEE Trans. Image Process*, vol. 33, pp. 1826–1837, 2024.

[10] P. R. G. Cadena, Y. Qian, C. Wang, and M. Yang, "SPADE-E2VID: Spatially-adaptive denormalization for event-based video reconstruction," *IEEE Trans. Image Process*, vol. 30, pp. 2488–2500, 2021.

[11] C. Ge, X. Fu, K. Wang, and Z.-J. Zha, "Event-based video reconstruction with deep spatial-frequency unfolding network," *IEEE Trans Image Process*, vol. 34, pp. 1779–1794, 2025.

[12] L. Zhu, Y. Zheng, Y. Zhang, X. Wang, L. Wang, and H. Huang, "Temporal residual guided diffusion framework for event-driven video reconstruction," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 411–427.

[13] W. Weng, Y. Zhang, and Z. Xiong, "Event-based video reconstruction using transformer," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 2543–2552.

[14] W. Ma, S. Ma, P. Meiresone, G. Allebosch, W. Philips, and J. Aelterman, "Event-based video reconstruction via attention-based recurrent network," *Neurocomputing*, vol. 632, 2025, Art. no. 129776.

[15] Y. Z. X. D. J. X. D. L. Z. T. Yangjie Cui and B. Gao, "WD-DETR: Wavelet denoising-enhanced real-time object detection transformer for robot perception with event cameras," 2025, *arXiv:2506.09098*.

[16] P. Shen and C. Jung, "Event-guided motion deblurring with wavelet-based cross-modal feature fusion," in *Proc. IEEE Int. Conf. Image Process.*, 2025, pp. 1876–1881.

[17] R. B. Sio-Hoi Ieng and E. Lehtonen, "Complexity analysis of iterative basis transformations applied to event-based signals," *Front. Neurosci*, vol. 12, 2018, Art. no. 373.

[18] C. Scheerlinck, N. Barnes, and R. Mahony, "Continuous-time intensity estimation using event cameras," in *Proc. Asian Conf. Comput. Vis.*, 2019, pp. 308–324.

[19] P. Bardow, A. J. Davison, and S. Leutenegger, "Simultaneous optical flow and intensity estimation from an event camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 884–892.

[20] J. Fang, Y. Dang, W. Zhao, B. Yu, Z. Wang, and J. Yin, "3DWSNet: A novel 3D wavelet spiking neural network for event-based action recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2025, pp. 11607–11613.

[21] Z. Cheng, X. Pan, X. Chen, and S. Fan, "Wavelet-based feature representation framework for event stream recognition," in *Proc. Int. Conf. Multimedia Expo*, 2025, pp. 1–6.

[22] R. Maqsood, P. Nunes, C. Conti, and L. D. Soares, "WaveE2VID: Frequency-aware event-based video reconstruction," in *Proc. IEEE Int. Conf. Image Process.*, 2025, pp. 570–575.

[23] R. Maqsood, P. Nunes, L. D. Soares, and C. Conti, "Efficient frequency-aware multiscale vision transformer for event-to-video reconstruction," in *Proc. 33rd Eur. Signal Process. Conf.*, 2025, pp. 606–610.

[24] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based optical flow using motion compensation," in *Proc. Eur. Conf. Comput. Vis.*, 2019, pp. 711–714.

[25] R. S. Stanković and B. J. Falkowski, "The haar wavelet transform: Its status and achievements," *Comput. Elect. Eng*, vol. 29, no. 1, pp. 25–44, 2003.

[26] W. Zou, M. Jiang, Y. Zhang, L. Chen, Z. Lu, and Y. Wu, "SDWNet: A straight dilated network with wavelet transformation for image deblurring," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2021, pp. 1895–1904.

[27] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-CNN for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 886–88609.

[28] H. Jiang, A. Luo, H. Fan, S. Han, and S. Liu, "Low-light image enhancement with wavelet-based diffusion models," *ACM Trans. Graph.*, vol. 42, no. 6, 2023, Art. no. 238.

[29] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 4003–4014.

[30] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[31] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[32] L. Sun et al., "Event-based fusion for motion deblurring with cross-modal attention," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 412–428.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations* 2015.

[34] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[35] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[36] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: An open event camera simulator," in *Proc. 2nd Conf. Robot Learn.*, 2018, pp. 969–982.

[37] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[38] E. Mueggler, H. Rebecq, G. Gallego, T. Delbrück, and D. Scaramuzza, "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM," *Int. J. Robot. Res*, vol. 36, pp. 142–149, 2016.

[39] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3D perception," *IEEE Robot. Automat. Lett.*, vol. 3, no. 3, pp. 2032–2039, Jul. 2018.