ISCTE IUL
REPOSITÓRIO
INSTITUTO UNIVERSITÁRIO DE LISBOA

Repositório ISCTE-IUL

Deposited in *Repositório ISCTE-IUL*:
2026-10-08

Deposited version:
Accepted Version

Peer-review status of attached file:
Peer-reviewed

Citation for published item:
Maqsood, R., Nunes, P., Conti, C. & Soares, L. D. (2025). WaveE2VID: Frequency-aware event-based video reconstruction. 2025 IEEE International Conference on Image Processing (ICIP), Proceedings (pp. 570?575). IEEE. https://doi.org/10.1109/ICIP55913.2025.11084548

Further information on publisher's website:
10.1109/ICIP55913.2025.11084548

Publisher's copyright statement:
This is the peer reviewed version of the following article: Maqsood, R., Nunes, P., Conti, C. & Soares, L. D. (2025). WaveE2VID: Frequency-aware event-based video reconstruction. 2025 IEEE International Conference on Image Processing (ICIP), Proceedings (pp. 570?575). IEEE. https://doi.org/10.1109/ICIP55913.2025.11084548, which has been published in final form at https://dx.doi.org/10.1109/ICIP55913.2025.11084548. This article may be used for non-commercial purposes in accordance with the Publisher's Terms and Conditions for self-archiving.

Use policy

Creative Commons CC BY 4.0
The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in the Repository
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Serviços de Informação e Documentação, Instituto Universitário de Lisboa (ISCTE-IUL)
Av. das Forças Armadas, Edifício II, 1649-026 Lisboa Portugal
Phone: +(351) 217 903 024 | e-mail: administrador.repositorio@iscte-iul.pt
https://repositorio.iscte-iul.pt

WaveE2VID: FREQUENCY-AWARE EVENT-BASED VIDEO RECONSTRUCTION

Ramna Maqsood, Paulo Nunes, Caroline Conti, and Luís Ducla Soares

Instituto de Telecomunicações, Instituto Universitário de Lisboa (ISCTE-IUL) Lisbon, Portugal

ABSTRACT

Event cameras, which detect local brightness changes instead of capturing full-frame images, offer high temporal resolution and low latency. Although existing convolutional neural networks (CNNs) and transformer-based methods for event-based video reconstruction have achieved impressive results, they suffer from high computational costs due to their linear operations. These methods often require 10M-30M parameters and inference times of 30-110 ms per forward pass at a resolution of 640×480 on modern GPUs. Furthermore, to reduce computational costs, these methods apply CNN-based downsampling, which leads to the loss of fine details. To address these challenges, we propose an efficient hybrid model, WaveE2VID, which combines the frequency-domain analysis of the wavelet transform with the spatio-temporal context modeling of a deep convolutional recurrent network. Our model achieves 50% faster inference speed and lower GPU memory usage than CNN and transformer-based methods, maintaining reconstruction performance on par with state-of-the-art approaches across benchmark datasets.

Index Terms— Event camera, wavelet transform, deep learning, video reconstruction.

1. INTRODUCTION

Event cameras, also known as neuromorphic cameras, offer a revolutionary approach to visual sensing by mimicking biological systems. Unlike conventional cameras that capture fixed-rate frames, event cameras operate asynchronously, detecting pixel-level light changes only when they exceed a pre-defined threshold. This novel mechanism provides significant advantages, including high temporal resolution, and low power consumption. However, the sparse and unstructured output of event cameras poses challenges for processing with traditional computer vision algorithms. Converting event streams into videos of sequential intensity images has become a key solution in event-based vision, enabling the application of frame-based algorithms to the reconstructed results for various vision tasks.

Deep learning techniques, particularly convolutional neural networks (CNNs), have shown remarkable success in event-based video reconstruction. E2VID, the first CNN-based model proposed by Rebecq *et al* [1], achieved a significant boost in reconstruction quality. Building on E2VID, Scheerlinck *et al* [2], introduced FireNet, a lightweight network that reduced inference time with only a minor compromise in accuracy. Subsequently, Stoffregen *et al* [3], showed reliance of these supervised methods on synthetic data generated by simulators like ESIM [4]. To address this, Federico *et al* [5], proposed the first self-supervised learning model, SSL-E2VID, which utilized optical flow and event-based photometric constancy to train networks without ground truth (GT). More recently, HyperE2VID [6] extended E2VID with hypernetworks [7], enabling dynamic parameterization for improved adaptability.

Despite these advancements, CNN-based methods face notable limitations. First, their high computational costs result in reconstruction times ranging from 30-110 milliseconds (ms) at a resolution of 640×480. To mitigate these costs, aggressive downsampling techniques are often employed to reduce input resolutions, which compromise fine details [8] in the reconstructed images. Furthermore, CNN's local receptive fields necessitate deep architectures to capture global dependencies [9], leading to increased computational overhead. Recently, ET-Net [10] a transformer-based model was introduced to capture global dependencies from event data. While promising, ET-Net suffers from high computational costs due to their linear operations required for modeling long-range dependencies, resulting in average inference times of 116.9 ms.

Frequency-aware analysis methods [11] have demonstrated their potential in event-based vision. The wavelet transform (WT), a well-established technique for multi-resolution and frequency-domain analysis, has recently gained attention in event-based vision for its integration into deep learning frameworks [12]. Event data is inherently sparse, dynamic, and rich in high-frequency information such as edges and brightness changes, which WT can efficiently decompose into localized, multi-scale subbands. By capturing both local and global contexts without requiring deep architectures, WT overcomes the limitations of convolutional methods [9]. Additionally, WT's ability to preserve fine-grained details while providing compact representations ensures computational efficiency [13]—a critical requirement for event-based vision tasks.

Building on these strengths, we propose WaveE2VID, a novel event-based video reconstruction model that integrates WT into a deep learning architecture. WaveE2VID combines the frequency-domain capabilities of WT with the spatio-temporal modeling strength of a deep convolutional recurrent network (DCRN) to achieve high-quality video reconstruction while significantly reducing computational overhead. WaveE2VID introduces two novel modules. Firstly, the frequency-aware downsampling module (FA-Encoder), utilizes 2D discrete wavelet transform (2D-DWT) for multi-resolution decomposition, effectively capturing global context and preserving frequency details. Unlike CNN-based downsampling methods that discard spatial information, the FA-Encoder retains critical features across diverse scales, ensuring efficient representations and a broader receptive field. Secondly, the

This work is funded by FCT/MECI through national funds and when applicable co-funded by EU funds under UID/50008: Instituto de Telecomunicações.

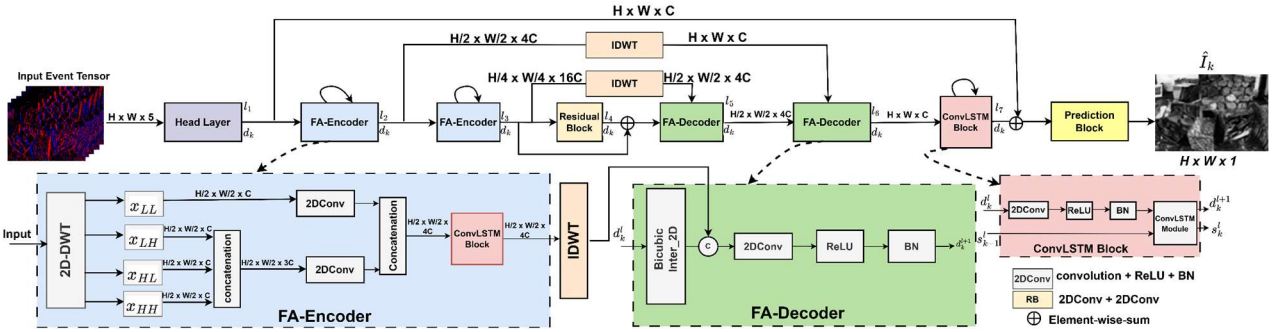


Figure 1: Overall proposed model pipeline for reconstruction of grayscale intensity images from event stream.

frequency-aware reconstruction module (FA-Decoder) leverages the inverse discrete WT (IDWT) to reconstruct high-fidelity intensity images by combining wavelet-based features with those learned by the DCRN, maintaining fine details and temporal consistency. To summarize the contributions:

1) We propose WaveE2VID, a novel hybrid event-based video reconstruction model that effectively balances reconstruction quality and efficiency, reducing parameters from 30M to 7M and inference time from 116.90 ms to 12.72 ms. This is achieved by combining frequency-domain representations from WT with the temporal modeling capability of DCRN.

2. RELATED WORK

2.1 Event-Based Video Reconstruction Methods

Video reconstruction is a prominent topic in event-based vision. Earlier approaches relied on non-machine learning techniques, like SLAM [14], regularization [15], and optimization [16] for reconstructing videos. The first machine learning approach, proposed by Barua *et al* [17], utilized a simulator to develop a sparse patch-based dictionary, enabling the matching of event patches to gradient patches. The success of deep learning, particularly CNNs, led to E2VID [1], which set new standards in reconstruction quality and spurred further advancements in models [6] [10]. Cadena *et al* [18] proposed SPADE-E2VID, a modified E2VID architecture by employing a new decoding scheme. One potential downside of these CNN-based works is their high computational cost, contrasting with the low-power and low-latency nature of event cameras. Therefore, we focus on the development of an efficient model that reconstructs good quality intensity images in resource constraint environments.

2.2 Wavelet Transform Methods

WT decomposes incoming signals into four subbands (i.e., one low-frequency and three high-frequency) using predefined functions, resulting in minimal information loss [9] as the transform is inherently designed for reversible compression (through IDWT). Finder *et al* [9], proposed a wavelet convolutional (WTConv) layer that increases the receptive field of convolutions effectively without suffering from over-parametrization. More closely related to our work, Liu *et al* [13] incorporate the WT within a modified UNet architecture [19] for down-sampling and up-sampling, demonstrating that the integration of CNNs and WT achieves superior performance in restoration tasks. Recently, Fang *et al* [12] proposed a spike wavelet transform (SWformer) model, which integrates WT into spiking neural networks (SNNs). Their work demonstrates that SWformer can efficiently and effectively extract both spatial and frequency features from event data, outperforming conventional SNNs in various tasks. In addition to this, other studies have investigated the application of WT in event-based vision, such as in fault diagnosis [20] and motion analysis [21] tasks. Despite this, to the best of our knowledge, no prior work has explored the potential of WT in event-based video reconstruction.

3. PROPOSED APPROACH

Our goal is to generate a sequence of intensity images $\{\hat{I}_k\}$ from an event stream. To achieve this, we begin with a preprocessing step that converts the event stream into voxel grids (Section 3.1) and each voxel grid is then processed by the proposed model. The model reconstructs a 2D grayscale image having dimensions $H \times W \times 1$ at each time step k, using the previous network state s_{k-1} and updating it to current state s_k. Here H and W represents the image height and width, respectively. Figure 1 shows the block diagram of our proposed approach.

3.1 Event Representation

To process the event stream, we follow a common approach in event-based literature [22]. Our network input is an event tensor $E(x, y, t)$ with dimensions $H \times W \times B$, where B is the number of temporal bins (referred to as the number of channels). To populate the event tensor, we employ trilinear voting (interpolation), where each event, (x_i, y_i, t_i, p_i), contributes its polarity, p_i, to its two closest temporal bins based on their normalized timestamps. Each event contribution to the event tensor E is determined using (1):

$$E(\mathrm{x}, \mathrm{y}, t_n) = \sum_i p_i \; max\,(0, 1 - |t_n - t_i^*|) \tag{1}$$

where t_n denotes the temporal bin index, p_i the polarity of the i^{th} event, indicating whether it corresponds to an increase or decrease in intensity, and t_i^* is the normalized version of the event timestamp, t_i, calculated using (2).

$$t_i^* = \frac{t_i - t_{min}}{t_{max} - t_{min}}(B - 1) \tag{2}$$

Here, t_{min} represents the earliest timestamp and t_{max} represents the latest timestamp observed in a section of the event data with N events. We set $B = 5$ and $N = 30000$ in our experiments.

3.2 Network Architecture

The proposed WaveE2VID architecture integrates convolutional layers (2DConv), convolutional long short-term memory (ConvLSTM) [23] and WT within a UNet-based structure [19]. A crucial aspect of the encoding scheme lies in the synergy between 2D-DWT layers and DCRN, while in the decoder, images are reconstructed using the detailed recovery of features by incorporating spatial details from IDWT. We describe each component of our proposed model in detail below.

Head Layer (HL): This layer employs a single 2DConv with a kernel size of 3×3, and stride of 1, to preserve spatial dimensions. The input event tensor, comprising 5 channels (i.e., bins), is transformed into a tensor of $C = 16$ channels.

Frequency-aware downsampling module (FA-Encoder): Our model consists of two FA-Encoder blocks. Each block works in 2D-DWT-CNN-ConvLSTM manner. The extracted features from HL are fed to the first FA-Encoder block, that initially apply 2D-DWT to decompose each channel into different frequency subbands. 2D-DWT yields coefficients corresponding to low-frequency components, which encapsulate essential structural details, and high-frequency components, which encode intricate fine details. The advantage of this approach lies in its cost-effectiveness, as expanding the channel dimensions which is usually accomplished through learnable convolutions, is achieved here using 2D-DWT with fixed kernels, making it a non-parametric method. To implement this, we employ the Haar [24] orthogonal wavelet functions due to their performance and faster computation on event data. The convolutional filters for Haar are defined in (3):

$$f_{LL}(x) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad f_{LH}(x) = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}$$
$$f_{HL}(x) = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \; f_{HH}(x) = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{3}$$

where f_{LL} filter serves the purpose of capturing the low-frequency information, while the f_{LH} , f_{HL} and f_{HH} filters are tasked with capturing high-frequency information in three distinct directions (horizontal, vertical and diagonal respectively) of input data. These four filters together form an orthogonal set that efficiently represents input data on a novel basis without redundancy. More specifically, a one-dimensional DWT is applied to the input, where low-pass and high-pass filters are convolved along the row and column directions separately. As a result, four subband images are obtained $(x_{LL}, x_{LH}, x_{HL}, x_{HH})$, which contain all the image information. According to this definition, the decomposition and convolution process are performed on input data using (4):

$$x_{LL} = (f_{LL} * x) \downarrow_2, \quad x_{LH} = (f_{LH} * x) \downarrow_2,$$
$$x_{HL} = (f_{HL} * x) \downarrow_2, \quad x_{HH} = (f_{HH} * x) \downarrow_2 \tag{4}$$

where $*$ is the convolution operation and $\downarrow_2$ denotes downsampling by a factor of 2. In conventional CNN, a single kernel is used for downsampling. In our approach, we employ both high-frequency and low-frequency filters as kernels for data decomposition. This dual-filter design enhances the inclusivity of input pixels, effectively expanding the receptive field [9]**.** Unlike state-of-the-art methods [1] [2] [3] [18]**,** which rely on 3×3 or 5×5 filters, our backbone integrates 2×2 low and high-frequency filters. By combining four such filters, we achieve an effective filter size equivalent to 8×8, with faster computation. Note that the receptive field only does not assure the precise reconstruction of data, other factors also contribute to the architecture's performance. Therefore, to enhance the module's capacity to extract the spatio-temporal context of data, the FA-Encoder incorporates a refinement step using DCRN. In this phase, the low-frequency subband, x_{LL}, undergoes a 3×3 2DConv followed by rectified linear unit (ReLU) activation and batch normalization (BN) layer to extract deeper structural features. The three high-frequency subbands are concatenated and processed through a separate 3×3 2DConv and ReLU activation, to extract intricate fine-grained details (shown in Figure 1). By processing each frequency subband independently, the module achieves a finely tuned adaptation to the intricate feature representations within each frequency band. To impose the long-range dependencies and maintain hidden states, we employ a ConvLSTM layer (details below). At the end of this stage, the down-sampled and frequency-enhanced features are reconstructed using IDWT, which is then passed to the FA-Decoder layer for further processing.

Residual Block (RB): Our architecture incorporates a single residual block to mitigate overfitting and preserve the network's performance. This block consists of two 2DConv layers, each with a kernel size of 3×3, designed to maintain the spatial and channel dimensions of the input. A skip connection adds the input features to the output of the second convolution before applying the ReLU.

Frequency-aware reconstruction module (FA-Decoder): The output dimensions from the RB differ from the original input dimensions, necessitating a reconstruction process to recover the full dimensions of the intensity image. To achieve this, we designed two FA-Decoder blocks, followed by a ConvLSTM block, to ensure accurate and detailed reconstruction. Each FA-Decoder block is specifically tailored to upscale the feature maps spatial resolution while preserving the channel consistency required to align with the FA-Encoder's output. Notably, these blocks benefit from the detailed recovery of features by incorporating spatial details from the IDWT (also shown in Section 4.4). Utilizing the wavelets orthogonal properties, our model allows for the recovery of down-sampled features using IDWT. The IDWT reconstructs the original image resolution decomposed and processed by the FA-Encoders. This reconstruction is concatenated with the reconstructed image via bicubic interpolation of FA-Decoder block, where the upsampling factor is 2 to maximally alleviate the adversarial effects. The interpolated features are then refined through a 2DConv layer with a 3×3 kernel size, allowing the decoder to capture local spatial relationships and further enhance feature representation (Figure 1).

ConvLSTM Block: Following [18], to mitigate flickering artifacts, we employ a ConvLSTM block that utilizes a 3×3 convolutional kernel with the same number of input and hidden layers as the previous layer. This layer produces the output that utilizes the previous states to enhance the temporal stability of reconstruction.

Prediction Block: This layer is a standard convolutional layer with a 1×1 kernel size and, unlike [1], we apply BN before the sigmoid activation function to this layer.

Loss Function: We used a combination of two loss functions for training. The first is the reconstruction loss $\mathcal{L}_k^R = d(\hat{I}_k, I_k)$, where d denotes the learned image patch similarity (LPIPS) distance [25]. The LPIPS metric measures perceptual similarity by passing both the reconstructed and target images through a VGG network [26], pre-trained on the ImageNet dataset [27]. Secondly, we use temporal consistency (TC) loss [28], as employed in [1], denoted as $\mathcal{L}_k^{TC}$, which aims to enforce temporal coherence between consecutive images based on the GT optical flow F_{k-1}^k and compute the photometric error. The temporal consistency loss is $\mathcal{L}_k^{TC} = M_{k-1}^k \left\| \hat{I}_k - W_{k-1}^k(\hat{I}_{k-1}) \right\|_1$ where W_{k-1}^k represents the result of warping the reconstruction at time step $k-1$ to time step k using the optical flow, and M_{k-1}^k is a weighting term that accounts for occlusions. The final loss, denoted as $\mathcal{L}$, is a weighted sum of both loss functions over L consecutive images and computed using (5) where λ_{TC} is the weight assigned to the temporal consistency loss (set to 5 empirically to balance with the reconstruction loss, as is also done in [1]). We train the recurrent network with a sequence length $L = 40$ and $L_0 = 2$.

$$\mathcal{L} = \sum_{k=1}^{L} \mathcal{L}_k^R + \lambda_{TC} \sum_{k=L_0}^{L} \mathcal{L}_k^{TC} \tag{5}$$

4. RESULTS AND EXPERIMENTAL SETUP

In this section, we first discuss the datasets used, followed by a

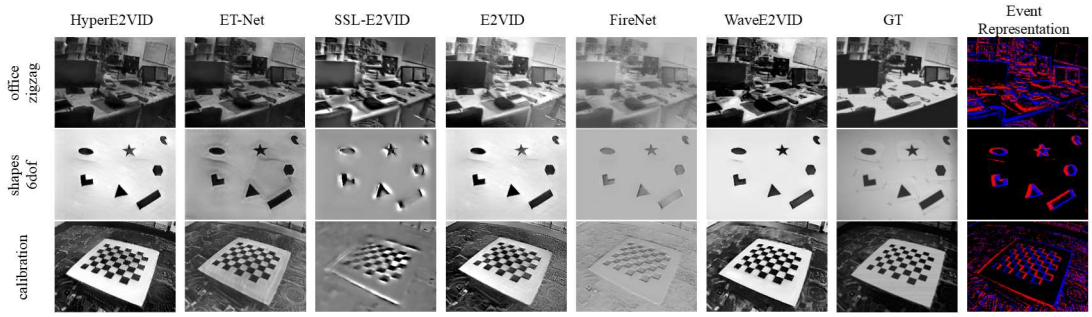


Figure 2: Visual results on ECD sequences with corresponding GT. Our approach can reconstruct good quality images.

Table 1: Comparison of computational costs.

Method	Parameters	Time (ms) 240 × 180	Time (ms) 640 × 480
E2VID, E2VID+ [1] [3]	10712090	6.2	30.88
FireNet,FireNet+ [2] [3]	**39999**	**1.6**	**2.03**
SPADE-E2VID [18]	12493030	16.1	26.34
ET-Net [10]	22184698	32.1	116.90
HyperE2VID [6]	10149881	6.6	18.82
WaveE2VID (Ours)	7432193	3.03	12.72
SPADE-WG (Ours)	12058089	12.81	20.05
E2VID-WG (Ours)	9596161	5.1	12.82

Table 2: Quantitative results on the ECD dataset.

Method	MSE	SSIM	LPIPS	TC
FireNet [2]	0.102	0.498	0.336	0.042
FireNet+ [3]	0.076	0.496	0.326	0.035
E2VID [1]	0.064	0.502	0.386	0.013
E2VID+ [3]	**0.045**	0.520	0.313	0.012
SSL-E2VID [5]	0.096	0.415	0.432	0.068
SPADE-E2VID [18]	0.091	0.462	0.387	0.048
ET-Net [10]	0.055	0.519	**0.235**	0.013
HyperE2VID [6]	0.054	**0.542**	0.265	0.014
WaveE2VID (Ours)	0.051	0.524	0.281	**0.011**

Table 3: BRISQUE score on the high speed and HDR dataset.

Method	hdr_sun	hdr_tunnel	hdr_selfie	Avg.
E2VID [1]	19.75	20.52	14.52	18.72
FireNet [2]	30.25	21.56	22.25	24.96
HyperE2VID [6]	**17.36**	19.75	**13.58**	16.89
WaveE2VID(Ours)	18.56	**14.65**	16.93	**16.71**

comparison of our approach with existing methods from two perspectives: i) computational performance, and ii) reconstruction accuracy. Lastly, we provided ablation study on proposed method.

4.1 Datasets

Training Dataset: Following the approach in [3], we synthesized the training dataset using the ESIM [4] rendering engine. Specifically, the "Multiple-Object-2D" simulation was employed, combining background images from the MSCOCO dataset [29] and foreground objects provided by [3]. The dataset consists of 280 sequences, each spanning 10 seconds, with event streams, GT intensity frames, and optical flow maps generated at an average rate of 51 Hz. Simulations were conducted with a sensor resolution of 240×180 pixels. For the contrast threshold settings, the values varied within each sequence between 0.1 and 1.5 in ascending order.

***Testing Datasets*:** We evaluated our model on two real-world datasets: i) Event camera dataset (ECD) [30], which is a widely recognized benchmark dataset used for evaluating event-based vision methods, comprising seven sequences recorded using DAVIS240C [31] sensor with a resolution of 240×180; and ii) The sequences from the high speed and HDR dataset [32], having a higher sensor resolution of 640×480, recorded by a Samsung DVS Gen3 [33] device.

Implementation Details: WaveE2VID is implemented in PyTorch, trained for 300 epochs with a batch size of 10, using Adam optimizer (10^{-4} learning rate, 10^{-3} weight decay). We applied data augmentation with random crops of 128×128 and flips (vertical/horizontal with 50% probability). Our model included noise, pause and hot-pixel augmentation as detailed in [3].

***Comparison Methods*:** We evaluated WaveE2VID using MSE (lower is better), SSIM (higher is better), and LPIPS (lower is better). We compared our model to eight state-of-the-art methods. As the GT images of ECD are darker, we follow a common approach [2] [18] of normalizing the images into the range of [0,1] to enable comparison with reconstructed images, without applying post-processing to the reconstructed images. Additionally, to evaluate the generalizability of our model, we developed wavelet-guided (WG) versions of E2VID and SPADE-E2VID, named E2VID-WG and SPADE-WG. In SPADE-WG, two ConvLSTM blocks were replaced by FA-Encoder blocks, and in E2VID-WG, both encoder and decoder layers were replaced with FA-Encoder and FA-Decoder blocks. For SPADE-WG, we adopted the many-to-one training loss strategy from the original work [18], to accelerate the training process. E2VID-WG, however, employed the same training loss function as WaveE2VID.

4.2 Computational Performance

In Table 1 we compare the computational cost of the WaveE2VID with existing methods. Bold values indicate the best-performing model, while underlined values represent the second-best. The comparison considers: i) the number of parameters, and ii) inference time in ms, evaluated at resolutions of common event cameras, such as DAVIS240C (240×180) and Samsung DVS Gen3 (640×480)**,** using an RTX 3080 GPU. WaveE2VID is 50% smaller than CNN-based models excluding FireNet (e.g., E2VID, SPADE-E2VID, HyperE2VID) and 80% smaller than the transformer-based ET-Net. While FireNet and FireNet+ are faster, they compromise reconstruction quality, as evidenced in Figure 2 .As shown in the last two rows of Table 1, our proposed WG models are also computationally more efficient than the original counterparts.

4.3 Reconstruction Accuracy

Quantitative results: Table 2 shows that WaveE2VID significantly outperforms methods such as SSL-E2VID, E2VID, FireNet, SPADE-E2VID, achieving up to 40% reduction in LPIPS while maintaining superior SSIM and MSE scores. These results highlight the advantage of integrating wavelet-based hybrid architecture into event-based video reconstruction. Although our model achieves comparable performance to advanced architectures such as HyperE2VID and ET-Net, these models leverage complex mechanisms, such as hypernetworks and global attention, which

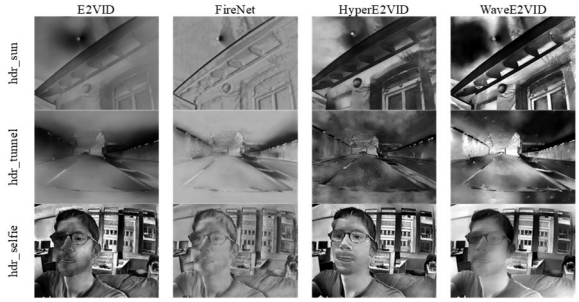


Figure 3: Visual results on the high speed and HDR dataset.

Table 4: Results of proposed WG models on ECD dataset

Method	MSE	SSIM	LPIPS
E2VID [1]	0.064	0.502	0.386
E2VID-WG (Ours)	**0.051**	**0.542**	**0.361**
SPADE-E2VID [18]	0.091	0.462	0.387
SPADE-WG (Ours)	**0.084**	**0.478**	**0.251**

contribute to edge in performance. However, our method stands out for its computational efficiency, making it more suitable for resource-constrained environments. Moreover, we compute TC loss [28] to measure the continuity of content between video frames, where lower values indicate best-performing model. Since the GT images for the high speed and HDR dataset [32] are not available, we used the no-reference Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [34]. Table 3 shows the results of WaveE2VID compared to three video reconstruction benchmarks, chosen for their computational efficiency. The lower BRISQUE values indicate that our model generalizes well on the high-resolution sensor. Furthermore, Table 4 demonstrates that our proposed WG models have outperformed the original models in terms of all metrics by a great margin.

Qualitative Results: Figure 2 presents a visual comparison of the proposed model on the ECD dataset. These visual results demonstrate that, despite requiring fewer computational resources, our method can also reconstruct details, comparable to resource-intensive approaches. Additionally, our model effectively reconstructs smooth regions with significantly fewer smearing artifacts compared to CNN-based methods. Visual results for temporal consistency in reconstructed videos can be seen in supplementary material at https://sigport.org/documents/wavee2vid-frequency-aware-event-based-video-reconstruction. Figure 3 shows results on the high speed and HDR dataset, demonstrating the generalizability of our method to a different sensor. To evaluate the reconstruction ability on this dataset, we reconstruct sequences using $N = 45000$ due to the higher sensor resolution (640×480 instead 240×180). The qualitative results of our proposed WG models on various event-based datasets (see supplementary material) also demonstrate the generalizability of our proposed. approach. Both models can effectively remove the artifacts.

Fast camera motion: We evaluated the reconstruction capabilities of WaveE2VID in comparison to larger networks, such as E2VID, under fast camera motion scenarios. For this assessment, we utilized the ECD_Fast subset proposed in [6]. As shown in Figure 4, despite having fewer parameters, our network can reconstruct good quality intensity images in challenging scenarios, comparable to those produced by larger networks.

4.4 Ablation Study

w/o FA-Encoder: We assess the impact of our FA-Encoder where our feature decomposition and extraction operations are replaced

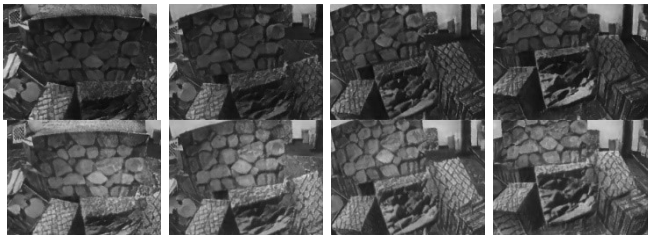

Figure 4: Comparison between E2VID (top row) and WaveE2VID (bottom row) in a challenging scenario.


Figure 5: Visual comparison of WaveE2VID (bottom row) and *w/o FA* (top row).

Table 5: Ablation results on different network settings.

Experiment Name	MSE	SSIM	LPIPS
w/o FA	0.068	0.502	0.378
w/o IDWT	0.061	**0.525**	0.321
WaveE2VID (ours)	**0.051**	0.524	**0.281**

with a simple convolution downsampling mechanism (stride=2). We train *w/o FA model* with our training dataset under the same settings as WaveE2VID Figure 5, presents a sequence of reconstructed images from the dynamic_6dof scene in the ECD dataset. The top row shows results from *w/o FA*, while the last row corresponds to WaveE2VID. A closer examination (highlighted by red boxes) reveals that WaveE2VID preserves fine details more effectively. Table 5 shows, the impact of this experiment on all metrics.

w/o IDWT: We assess the contribution of IDWT in the reconstruction phase of our WaveE2VID model. In this experiment, we removed the IDWT layers while retaining all other modules, including the skip connections, to isolate the impact of IDWT. As shown in Table 5, the model without IDWT still demonstrates good generalization in terms of SSIM scores. However, notable differences were observed in the LPIPS and MSE metrics, indicating that the absence of IDWT affects the model's ability.

5. CONCLUSION

In this work, we proposed WaveE2VID, a novel hybrid model that integrates traditional frequency-aware wavelet transforms (WT) with deep neural networks, effectively combining the strengths of both approaches. Our design emphasizes the need for smaller, efficient architectures that strike a balance between accuracy and efficiency, leveraging the low-latency and low-power characteristics of event data. Additionally, we introduced SPADE-WG and E2VID-WG, which enhance the reconstruction quality of their base models while maintaining lower computational overhead. Our findings demonstrate that wavelet-based hybrid architecture provides effective solutions to computational challenges in event-based vision. We suggest that integrating WT with more complex architectures could yield further improvements, to be explored in future work. We believe this work offers valuable insights for the community in designing efficient and frequency-aware architectures for event-based video reconstruction.

REFERENCES

[1] H. Rebecq, R. Ranftl, V. Koltun and D. Scaramuzza, "High Speed and High Dynamic Range Video with an Event Camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 43, no. 6, pp. 1964-1980, 2021.

[2] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. E. Mahony and D. Scaramuzza, "Fast Image Reconstruction with an Event Camera," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 156-163.

[3] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, Drummond, B. Tom, K. Nick, M. Lindsay and R. Mahony, "Reducing the Sim-to-real Gap for Event Cameras," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 534-549.

[4] H. Rebecq, D. Gehrig and D. Scaramuzza, "ESIM: An Open Event Camera Simulator," in *Conference on Robot Learning PMLR*, 2018, pp. 969-982.

[5] F. Paredes-Vallés and G. C. H. E. de Croon, "Back to Event Basics: Self-Supervised Learning of Image Reconstruction for Event Cameras via Photometric Constancy," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3445-3454.

[6] B. Ercan, O. Eker, C. Saglam, A. Erdem and E. Erdem, "HyperE2VID: Improving Event-Based Video Reconstruction via Hypernetworks," *IEEE Transactions on Image Processing,* vol. 33, pp. 1826-1837, 2024.

[7] D. Ha , A. Dai and Q. V. Le, "Hypernetworks," in *arXiv preprint arXiv:1609.09106*, 2016.

[8] R. Sunkara and T. Luo, "No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects," in *Machine Learning and Knowledge Discovery in Databases*, 2023, pp. 443-459.

[9] S. E. Finder, R. Amoyal, E. Treister and O. Freifeld, "Wavelet Convolutions for Large Receptive Fields," in *European Conference on Computer Vision(ECCV)*, 2025, pp. 363-380.

[10] W. Weng, Y. Zhang and Z. Xiong, "Event-based Video Reconstruction using Transformer," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2563-2572.

[11] T. Kim, C. Hoonhee and Y. Kuk-Jin, "Frequency-aware Event-based Video Deblurring for Real-World Motion Blur," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24966-24976.

[12] Y. Fang, Z. Wang, l. Zhang, J. Cao, H. Chen and R. Xu , "Spiking Wavelet Transformer," in *European Conference on Computer Vision (ECCV)*, 2025, pp. 19-37.

[13] P. Liu, H. Zhang, K. Zhang, L. Lin and W. Zuo, "Multi-level Wavelet-CNN for Image Restoration," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 773-782.

[14] H. Rebecq, T. Horstschaefer, G. Gallego and D. Scaramuzza, "EVO: A Geometric Approach to Event-Based 6-DOF Parallel Tracking and Mapping in Real Time," *IEEE Robotics and Automation Letters,* vol. 2, no. 2, pp. 593-600, 2017.

[15] G. Munda, C. Reinbacher and T. Pock, "Real-Time Intensity-Image Reconstruction for Event Cameras Using Manifold Regularisation," *International Journal of Computer Vision,* vol. 126, no. 12, pp. 1381-1393, 2018.

[16] P. Bardow, A. J. Davison and S. Leutenegger, "Simultaneous Optical Flow and Intensity Estimation from an Event Camera," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 884-892.

[17] S. Barua, Y. Miyatani and A. Veeraraghavan, "Direct Face Detection and Video Reconstruction from Event Cameras," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1-9.

[18] P. R. G. Cadena, Y. Qian, C. Wang and M. Yang, "SPADE-E2VID: Spatially-Adaptive Denormalization for Event-Based Video Reconstruction," *IEEE Transactions on Image Processing,* vol. 30, pp. 2488-2500, 2021.

[19] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234-241.

[20] R. Guang, X. Li, Y. Lei, N. Li and B. Yang, "Non-Contact Machine Vibration Sensing and Fault Diagnosis Method Based on Event Camera," in *Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS)*, 2023, pp. 1-5.

[21] G. Chen, L. Hong, J. Dong, P. Liu, J. Conradt and A. Knoll, "EDDD: Event-Based Drowsiness Driving Detection Through Facial Motion Analysis With Neuromorphic Vision Sensor," *IEEE Sensors Journal,* vol. 20, no. 11, pp. 6170-6181, 2020.

[22] A. Zihao Zhu, L. Yuan, K. Chaney and K. Daniilidis, "Unsupervised Event-based Optical Flow using Motion Compensation," in *European Conference on Computer Vision (ECCV) Workshops*, Springer, 2018, pp. 711-714.

[23] 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," *Advances in Neural Information Processing Systems,* vol. 28, pp. 2003-2103, 2015.

[24] R. S. Stankovic and B. J. Falkowski, "The Haar Wavelet Transform: its status and achievement," *Computers & Electrical Engineering,* vol. 29, no. 1, pp. 25-44, 2003.

[25] R. Zhang, P. isola, A. A. Efros, E. Shechtman and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586-595.

[26] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *3rd International Conference on Learning Representations (ICLR)*, 2015.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein and et al, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision,* vol. 115, no. 3, pp. 211-252, 2015.

[28] W.-S. Lei, J. B. Huang, O. Wang, E. Shechtman, E. Yumer and M. H. Yang, "Learning Blind Video Temporal Consistency," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 170-185.

[29] T.-Y. Lin, M. Marie, S. Belongie, , J. Hays, P. Perona, D. Ramanan, P. Dollar and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 740-755.

[30] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck and D. Scaramuzza, "The Event-Camera Dataset and Simulator: Event-based data for Pose Estimation, Visual Odometry, and SLAM," *The International Journal of Robotics Research,* vol. 36, no. 2, pp. 142-149, 2017.

[31] C. Brandli, R. Berner, M. Yang, S.-C. Liu and T. Delbruck, "A 240 × 180 130 dB 3 µs Latency Global Shutter Spatiotemporal Vision Sensor," *IEEE Journal of Solid-State Circuits,* vol. 49, no. 10, pp. 2333-2341, 2014.

[32] H. Rebecq, R. Ranftl, V. Koltun and D. Scaramuzza, "High Speed and High Dynamic Range Video with an Event Camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 43, no. 9, pp. 1964-1980, 2019.

[33] B. Son, Y. Suh, S. Kim, H. Jung, J.-S. Kim, C. Shin, K. Park, K. Lee, J. Park, J. Woo, Y. Roh, H. Lee, Y. Wang, l. Ovsiannikov and H. Ryu, "4.1 A 640×480 Dynamic Vision Sensor with a 9µm Pixel and 300Meps Address-Event Representation," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2017, pp. 66-67.

[34] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *IEEE Transactions on Image Processing,* vol. 21, no. 12, pp. 4695-4708, 2012.