Efficient Frequency-Aware Multiscale Vision Transformer for Event-to-Video Reconstruction

Ramna Maqsood, Paulo Nunes, Luís Ducla Soares, and Caroline Conti
Instituto de Telecomunicações, Instituto Universitário de Lisboa (ISCTE-IUL) Lisbon, Portugal
{ramna.maqsood, paulo.nunes, lds, caroline.conti}@lx.it.pt

***Abstract*—Event-to-video (E2V) reconstruction is a critical task in event-based vision, benefiting from the advantages of event cameras, such as high dynamic range and low latency. However, existing deep learning reconstruction methods often prioritize temporal consistency and over-emphasize low-frequency features, leading to blur artifacts and loss of fine details. To overcome these limitations, we propose a novel frequency-aware multiscale vision transformer model for E2V reconstruction (MSViT-E2V). Our model employs wavelet-based decomposition to extract features at multiple scales, preserving fine-grained details through multi-level wavelet-based downsampling blocks, followed by transformer blocks for multiscale feature aggregation and long-range dependency modeling. Extensive experiments on various event datasets demonstrate that our model not only minimizes artifacts and preserves fine details but also reduces computational costs by up to 50% compared to the transformer-based model ET-Net.**

***Index Terms*—Event-based vision, frequency-domain analysis, video reconstruction, vision transformer.**

I. INTRODUCTION

Event-to-video (E2V) reconstruction has emerged as a highly promising research area, driven by the unique advantages of event cameras over traditional frame-based cameras, such as high dynamic range and low power consumption. Early approaches to E2V reconstruction primarily relied on non-machine learning techniques [1]. However, the introduction of convolutional neural networks (CNNs) by Rebecq *et al* [2], with their E2VID model, marked a significant breakthrough, achieving state-of-the-art (SOTA) results and paving the way for CNN-based models [3]–[6]. Scheerlinck *et al* [4] introduced FireNet, which employs a lighter network to achieve faster E2V reconstruction speed. Stoffregen *et al*, [3] proposed a pipeline for synthetic dataset generation and proposed enhanced versions of E2VID and FireNet, trained on synthetic data: E2VID+ and FireNet+. Recently, Ercan *et al* [5] introduced HyperE2VID, a dynamic architecture using hypernetworks for adaptive inference. These models face inherent limitations due to the local receptive fields of convolutional kernels, which struggle to capture long-range dependencies. This limitation is particularly critical in E2V, where understanding spatial dependencies and processing complex texture patterns is crucial. Furthermore, CNN-based methods often fail to handle structures with significant internal variations in texture and shape [7]. These limitations have recently been addressed through the use of the multiscale transformer model ET-Net [7]. This model decomposes event data into multiple resolutions and employs a self-attention mechanism to aggregate global features across different scales. Despite its improved performance, ET-Net still exhibits some limitations. For instance, similar to CNN-based methods, it focuses on temporal consistency in continuous event streams leading to an over-reliance on low-frequency (LF) texture features [8], resulting in artifacts such as blur, over-smoothing, and the loss of fine details. Additionally, the ET-Net model is computationally expensive, which does not favor the low-latency nature of event cameras.

In event-based vision, preserving high-frequency (HF) features is crucial [8], [9]. Therefore, there is a pressing need for an innovative approach that can effectively preserve fine-grained details while maintaining computational efficiency. Frequency-aware methods have shown remarkable potential in event-based tasks [9]–[11]. Techniques like discrete wavelet transform (DWT), which excel in multi-resolution and frequency domain analysis, have been integrated into deep learning frameworks to better preserve fine details from event data [11]. Event data is inherently sparse and rich in HF information which DWT can efficiently decompose into multiscale subbands (one LF and three HF subbands). For example, [9] proposed a day-to-night event translation method using wavelet decomposition, demonstrating its ability to accurately preserve HF details such as edges. Similarly, a 3D DWT [10] is integrated into spike neural networks (SNNs) to decompose event data into LF and HF components at various scales. Fang *et al* [11] introduced a spiking wavelet transform (WT) that integrates DWT into SNNs. Their method effectively extracts spatial and frequency features from event data, outperforming traditional SNNs in various tasks. In [12], the author demonstrates the computational efficiency of linear basis transformations, such as DWT, on event data, when compared with deep learning models. While WT-based methods combined with deep learning have been explored for various event-based vision tasks, their potential for E2V reconstruction remains largely unexplored, to the best of our knowledge. Therefore, in this paper, we propose a novel efficient frequency-aware model that addresses the limitations of existing methods: multiscale vision transformer for E2V reconstruction (MSViT-E2V). Our model employs a UNet structure [13] to apply a multiscale feature aggregation strategy, where input features are first extracted at multiple scales using multi-level wavelet-based

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

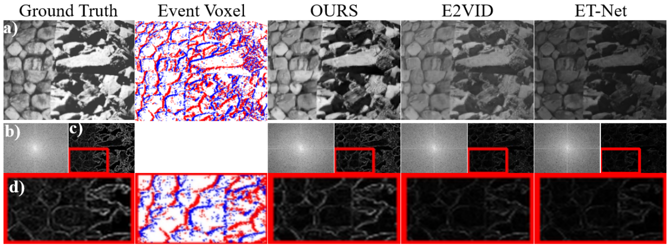

Fig. 1. Event-to-video reconstruction: a) In the first row, we have ground truth image with corresponding event voxel and 2D reconstructed intensity images, b) Fourier spectrum of intensity images, c) All three HF components obtained via DWT on intensity images, and d) zoom view (scaled for better visualization). While the event voxel resembles the HF map of the scene, reconstruction from E2VID and ET-Net remains blurry and lacks fine details.

downsampling blocks (WDBs) to process both LF and HF information. These multiscale features are then passed through transformer blocks (TBs) to model long-range dependencies, enabling robust feature extraction across different scales. Lastly, to adaptively weight features across spatial locations and channels, attention-based [14] upsampling blocks (UBs) are employed to reconstruct grayscale images at the original resolution. Our contributions can be summarized as:

- A novel frequency-aware multiscale vision transformer (MSViT-E2V) model is proposed that leverages wavelet-based decomposition and reconstruction to preserve fine and structural details in event data at multiple scales.
- The proposed framework achieves superior reconstruction quality and significant computational efficiency, compared to the existing SOTA ET-Net method.
- Extensive experiments demonstrate that the proposed solution is effective in overcoming issues like artifacts and loss of details in reconstructed videos (Fig 1).

II. Proposed Approach

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 [15], that converts the event stream into voxel grids, each voxel grid is then passed to our proposed model that reconstructs a 2D grayscale image having dimensions $H \times W \times 1$, where H and W represent the image height and width, respectively.

A. Event Representation

Given an event stream $\{e_i\}$, where $e_i = (x_i, y_i, t_i, p_i)$ represents the i-th event with spatial coordinates (x_i, y_i), timestamp $t_i \in [0, T]$, and polarity p_i (+1 for positive and -1 for negative polarity), we group events into intervals defined by consecutive ground truth (GT) image timestamps. Specifically, the k-th event group $E_k = \{e_i \mid T_{k-1} \leq t_i < T_k\}$ where T_k is the ending timestamp of the k-th group and $\Delta T = T_k - T_{k-1}$ is its duration. Following a common practice [2], we divide each group into five temporal bins ($B = 5$) as it captures temporal dynamics and maintain computational efficiency. The event timestamps in E_k are normalized to the range $[0, B-1]$ using: $t_i^* = \frac{t_i - T_{k-1}}{\Delta T}(B-1)$. Each event, e_i, contributes its polarity to the two closest bins via bilinear interpolation, forming a voxel grid, $V_k \in \mathbb{R}^{H \times W \times B}$, for the k-th group using (1), where t_n represents the temporal bin index.

$$V_k(x, y, t_n) = \sum_i p_i \max(0, 1 - |t_n - t_i^*|) \qquad (1)$$

B. MSViT-E2V

Fig. 2 shows the proposed reconstruction model that integrates WDBs, TBs, and attention-based UBs to jointly produce multi-resolution features, global contextual information, and spatial and channel-aware feature representations. The input event voxel, V_k, is first processed by a 3×3 2D convolutional layer (2DConv), transforming it into a feature representation, $X \in \mathbb{R}^{H \times W \times C}$, where C is input channels, i.e., $C = 32$.

Wavelet-based Downsampling Blocks (WDBs): Three WDBs process the input in the frequency domain, aiming to overcome spatial information loss and limited receptive fields in existing CNN-based methods [3]–[6]. Specifically, the WDB decomposes and downsamples the input feature maps, X, using 2D-DWT into four subbands: $\{X_{LL}, X_{LH}, X_{HL}, X_{HH}\} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$, where X_{LL} is the LF subband, containing the structural information of the input, and $\{X_{LH}, X_{HL}, X_{HH}\}$ are the three HF subbands, representing fine details in the horizontal, vertical, and diagonal directions, respectively. The decomposition is achieved by applying 1D-DWT along the rows and columns of the input, using low and high-pass filters. We employ the Haar wavelet function due to its computational efficiency [16]. To fully exploit the frequency decomposition capabilities of the DWT, the subbands are processed independently. X_{LL} is then passed through a 3×3 depth-wise convolution (DWConv) to refine the structural information while preserving spatial resolution. Meanwhile, the HF subbands are concatenated along the channel dimension and processed through another 3×3 DWConv to capture fine details. Each subsequent WDB applies the same wavelet decomposition process, but instead of operating on the original input features, it takes the LF subband from the preceding WDB, X_{LL}^{i-1}, as input, where $i \in \{1, 2, 3\}$. This multi-level decomposition further refines structural details at progressively lower resolutions, ensuring a more effective hierarchical representation. The wavelet decomposition increase. Therefore, to maintain memory efficiency, the processed LF and HF subbands are first concatenated along the channel dimension into X_{concat}^i (see Fig. 2a) and then a 3×3 2DConv is used to reduce the channel dimensions and extract underlying features. To capture temporal data dependencies, a convolutional long-short-term memory (ConvLSTM) module [17] is employed, where the 1×1 2DConv refines feature representations before feeding them into the ConvLSTM for temporal modeling and producing the output, X_{out}^i. Within our WDBs, we perform an intermediate reconstruction that acts as a skip connection between the WDB and the UBs. Specifically, the concatenated feature map, X_{concat}^i, undergoes inverse WT (IWT) to restore spatial resolution while preserving both LF and HF details. This reconstruction serves as an intermediate step rather than the final model output as shown in Fig. 2a.

Transformer Blocks (TBs): As shown in Fig. 2b, each TB processes multi-resolution feature maps of shape (H, W, C) generated by each WDBs. The TB reshapes the input into a

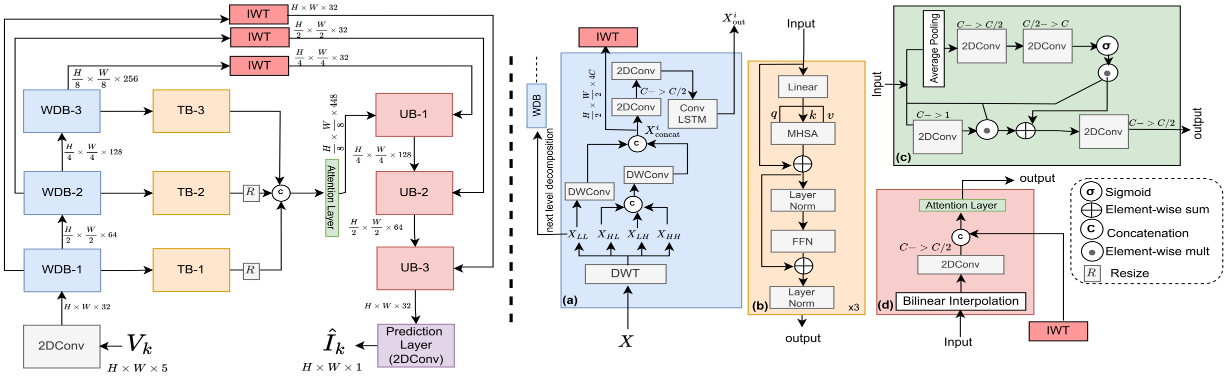

Fig. 2. Overall proposed model (Left side). On the right side, we show detailed working of (a) WDBs, (b) TBs, (c) attention layer, and (d) UBs.

sequence of tokens, $X^i_{\text{reshaped}} \in \mathbb{R}^{N \times C}$, where $N = H \times W$ and projects them into query (q), key (k), and value (v) using learnable weight matrices. These projections are passed to the multi-head self-attention (MHSA) [18] mechanism, which computes attention scores to capture global dependencies across the spatial dimensions. The output of the MHSA is combined with the original input via a residual skip connection, followed by layer normalization (LN) to stabilize training. The normalized features are then passed through a feed-forward network (FFN), consisting of two linear transformations with a Gaussian error linear unit activation and a dropout (0.1 in our experiments) in between, to refine the feature representations. Finally, another residual skip connection and LN are applied to produce the output of the TB. After processing the input at different scales, the outputs of all three TBs are resized to a common resolution of 16×16 via bilinear interpolation. The resized feature maps are concatenated along the channel dimension.

Attention Layer: To perform local refinement, by re-weighting the concatenated multi-resolution features across both spatial and channel dimensions, we employ an attention layer, inspired by [14]. This enables our model to focus on the most informative regions by adjusting the features in both channel and spatial domains. In the channel domain, the concatenated output of the three TBs is first subjected to average pooling to squeeze the features, followed by two 1×1 2DConv layers and a sigmoid activation to generate a channel attention map. Simultaneously the input is processed by a 1×1 2DConv to produce a spatial attention map. Both maps are multiplied with their respective input features to refine the feature representations (see Fig. 2c). Finally, another 3×3 2DConv layer is applied to reduce the channel dimensions by half, ensuring efficient feature extraction.

Upsampling Blocks (UBs): Unlike in [2]–[7], we adopt attention-based UBs to improve image reconstruction (see Fig. 2d). Each UB begins with bilinear interpolation, increasing the spatial resolution by a factor of 2, followed by a 3×3 2DConv to reduce channel dimensions while preserving spatial details. To leverage multi-resolution features, the upsampled output is concatenated with the IWT of the corresponding scale, where the IWT acts as a skip connection between the WDBs and the UBs. This ensures the preservation of both LF structures and HF details. However, the concatenated features may contain redundant or noisy information, which can cause ghosting to appear in the reconstructed images. To address this, we introduce an attention layer (see details in Fig. 2c) in UBs to adjust the features in both spatial and channel domains. This refinement allows the network to focus on significant textures and structural details while suppressing noise.

Prediction Layer: This layer consists of a standard 1×1 2DConv layer, followed by batch normalization (BN) and a sigmoid activation function. This layer generates the final 2D grayscale intensity image $\hat{I}_k \in \mathbb{R}^{H \times W \times 1}$.

Loss Functions: For training, we used learned image patch similarity (LPIPS) and temporal consistency (TC) loss functions, as in [2]. LPIPS is a perceptual loss function that measures the similarity between the GT and the reconstructed image, whereas TC measures transition smoothness between consecutive video frames, penalizing discrepancies in motion or textures. The final loss, $\mathcal{L}$, is a weighted sum of both loss functions over L consecutive images, computed using (2):

$$\mathcal{L} = \sum_{k=1}^{L} \mathcal{L}^R_k + \lambda_{TC} \sum_{k=L_0}^{L} \mathcal{L}^{TC}_k \tag{2}$$

where $\mathcal{L}^R_k$ and $\mathcal{L}^{TC}_k$ are LPIPS and TC loss values at time k, λ_{TC} is the TC loss weight (set to 5 empirically to balance the reconstruction loss, as in [2]), and we train the recurrent network with a sequence length $L = 40$ and $L_0 = 2$.

III. Experimental Setup and Results

This section outlines the training and testing datasets, results, and network analysis of the proposed model.

Training Dataset: We follow a common approach to generate a synthetic training dataset [3], using the ”Multiple-Object-2D” rendering engine of ESIM [19]. This engine simulates multiple foreground objects moving across a background image, employing various 2D motion properties. Background images were selected from the MSCOCO dataset [20], while foreground objects were sourced from [3]. 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, produced at a rate of 51Hz, both captured at the resolution of 256×256 pixels. Note that our model can adapt to varying resolutions

TABLE I
QUANTITATIVE RESULTS ON **ECD**, **HQF**, AND **ECD_FAST** DATASETS. BEST IS IN **BOLD**, SECOND BEST IS <u>UNDERLINED</u>.

Methods	**ECD**			**HQF**			**ECD_fast**		
	MSE↓	SSIM ↑	LPIPS ↓	MSE ↓	SSIM ↑	LPIPS ↓	MSE ↓	SSIM ↑	LPIPS↓
FireNet	0.142	0.478	0.336	0.094	0.423	0.441	0.131	0.444	0.367
FireNet+	0.116	0.491	0.416	0.080	0.471	0.314	0.049	0.449	0.329
SSL-E2VID	0.096	0.385	0.442	0.082	0.421	0.467	0.109	0.389	0.415
SPADE-E2VID	0.101	0.442	0.397	0.077	0.400	0.502	0.049	0.449	0.329
E2VID	0.102	0.476	0.416	0.098	0.468	0.371	0.203	0.374	0.413
HyperE2VID	0.062	0.492	0.370	0.058	0.460	0.370	0.047	0.495	0.307
E2VID+	0.071	0.501	<u>0.286</u>	<u>0.036</u>	0.533	<u>0.252</u>	0.069	0.501	0.388
ET-Net	<u>0.065</u>	<u>0.523</u>	0.263	0.057	<u>0.483</u>	0.293	<u>0.057</u>	<u>0.532</u>	<u>0.354</u>
MSViT-E2V (Ours)	**0.048**	**0.561**	**0.224**	**0.059**	**0.555**	**0.241**	**0.046**	**0.582**	**0.302**

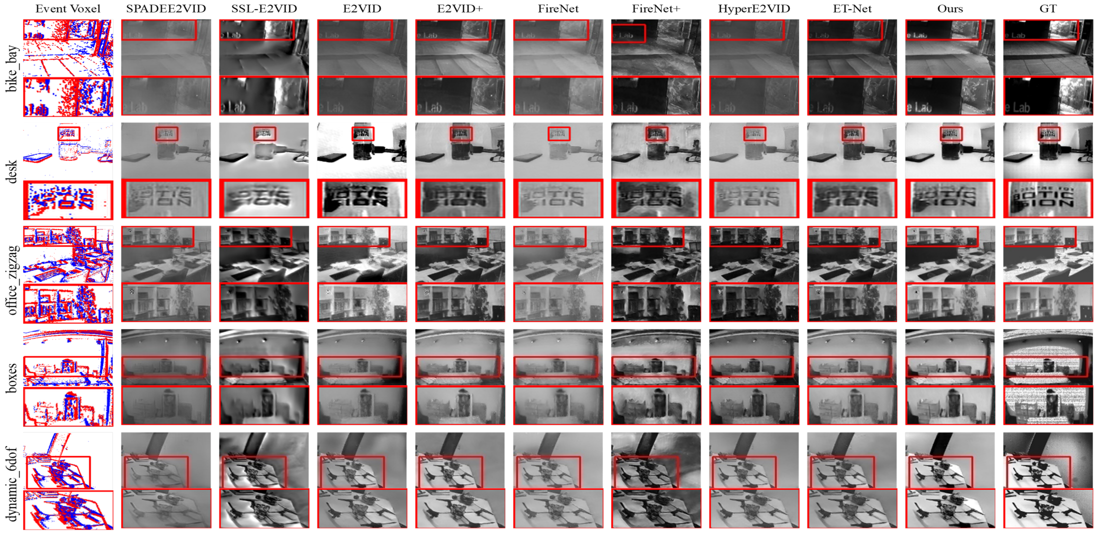

Fig. 3. Visual results on HQF (rows 1 and 2), ECD (rows 3 and 4), and ECD_fast (last row). We provided a magnified view of each reconstructed scene where visual differences are more prominent. Overall, our reconstructed scenes are more close to GT images in terms of contrast and fine details.

due to the use of fully convolutional operations, which process input in a resolution-agnostic manner.

Testing Datasets: To evaluate the performance of the proposed model, we utilized the two following benchmark datasets:

- The event camera dataset (ECD) [21] recorded with a sensor resolution of 240 × 180. This dataset features sequences from seven indoor environments. Within this dataset, we also evaluate our model on the ECD_fast subset introduced by [5], which consists exclusively of the fast camera motion sequences from the ECD.
- The high quality frame (HQF) dataset [3], recorded with the same resolution as ECD. This dataset comprises 14 sequences with a wider range of motion and scene types.

Evaluation Metrics: To evaluate the proposed model the following commonly used full reference metrics were used: mean square error (MSE) [↓], structural similarity index matrix (SSIM) [↑], and LPIPS [↓].

Implementation Details: The proposed model is implemented in PyTorch, trained for 300 epochs with a batch size of 4 on an RTX 3080 GPU. We use the Adam optimizer with an initial learning rate of 0.001, decaying by 10% every 50 epochs. Data augmentation includes random crop (128 × 128) and flip (probability 0.5), along with noise, pause, and hot-pixel augmentation as also done in [3].

Comparison with SOTA Methods: We compare our proposed method with eight SOTA methods for which publicly available code exists: FireNet [4], FireNet+ [3], E2VID [2], E2VID+ [3], SPADE-E2VID [6], SSL-E2VID [22], ET-Net [7], and HyperE2VID [5]. For fair evaluation, we re-trained all models on our synthetic dataset using the same settings as MSViT-E2V, except for E2VID and FireNet, which use pre-trained weights due to their original training on a different dataset. As the GT images of ECD are darker, we follow a common approach [2] of normalizing the images into the range of $[0, 1]$ to enable comparison with reconstructed images, without applying post-processing to the reconstructed images.

Table I shows the quantitative results achieved, in terms of average metrics, for each testing dataset. As can be seen in this table, the proposed approach achieved SOTA results for almost all metrics and datasets. As illustrated in Fig. 3, for the HQF dataset, our model preserves fine structural details and perceptually realistic results more effectively, which is also reflected in the SSIM and LPIPS metrics. Our SSIM scores are 7.27%, 4.13%, and 9.40% higher on the ECD, HQF, and ECD_Fast datasets, over the second-best method, demonstrating superior structural similarity and detail preservation. Furthermore, in challenging scenarios such as fast camera motion (ECD_fast), our model achieved significant gains across all metrics. Fig. 3 shows GT images in the rightmost column for comparison, highlighting the structural details preserved in the reconstructed images by each method. As evidenced in rows 1 and 2, our model preserves texture details more clearly than the other methods. Moreover, our method exhibits less artifacts in the reconstructed images compared with SOTA methods often

TABLE II
ABLATION RESULTS ON ECD AND HQF DATASETS.

Methods	X_{LL}	X_{LH}, X_{HL}, X_{HH}	Spatial-Branch	IWT	ECD		HQF	
					LPIPS↓	SSIM↑	LPIPS↓	SSIM↑
only X_{LL}	✓	✗	✗	✗	0.328	0.531	0.341	0.495
only X_{LL}, X_{LH}, X_{HL}	✓	✓	✗	✗	0.262	0.542	0.291	0.561
All Freq. subbands	✓	✓	✗	✓	0.231	0.558	0.262	0.566
w/o Freq.	✗	✗	✓	✓	0.304	0.452	0.325	0.499
Ours	✓	✓	✓	✓	**0.224**	**0.561**	**0.241**	**0.555**

containing blur and ghosting artifacts (see rows 3, 4, and 5). For more visual results, please refer to our GitHub repository.

Network Analysis: To analyze the impact of frequency components, we evaluated our model using various frequency selection strategies. All the models are trained under the same settings and training dataset as our MSViT-E2V. Table II reveals that using only X_{LL} subband provides reasonably adequate SSIM but results in a high LPIPS, indicating blurry reconstructions due to missing HF details for both datasets. Only X_{LL}, X_{LH} and X_{HL} model enhances perceptual sharpness but reduces SSIM, showing that edge details alone are insufficient for structural consistency. Incorporating all subbands significantly improves both perceptual quality and structural details, highlighting the importance of X_{HH} components. This aligns with the findings in [9] that X_{HH} is critical for transferring event-style characteristics, such as natural noise, making it essential for realistic E2V reconstructions. Notably, using a spatial branch (stride-based downsampling where $s = 2$) without frequency components degrades both metrics, confirming that frequency decomposition aids effective feature representation. Note that in this experiment we replace our WDBs with the recurrent convolutional blocks of ET-Net [7]. Finally, integrating IWT further refines results by incorporating multi-resolution features as skip connections, demonstrating its role in balancing sharpness and stability.

Computational Costs: Table III presents the computational cost of the proposed method at a resolution of 240×180, with inference times measured in milliseconds on an RTX 3080 GPU. Notice that our transformer-based model is almost 50% smaller than ET-Net, demonstrating that our model provides a good trade-off between accuracy and efficiency.

IV. CONCLUSION

In this paper, we proposed MSViT-E2V, a novel model for E2V reconstruction. It integrates wavelet-based decomposition blocks to extract multi-resolution features, transformer blocks to model global contexts, and attention-based upsampling blocks to minimize artifacts by focusing on important regions. Experiments on various event datasets show that MSViT-E2V reduces artifacts and restores fine details more efficiently than existing methods. Moreover, our model is almost 50% smaller than SOTA ET-Net, making it more computationally efficient. In the future, we plan to explore pure frequency-based transformers to further enhance performance and efficiency.

REFERENCES

[1] 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 Robot. Autom. Lett*, vol. 2, no. 2, pp. 593–600, 2017.

[2] 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, 2021.

TABLE III
COMPUTATIONAL COMPARISON WITH EXISTING METHODS.

Method	No of parameters	Inference time
E2VID, E2VID+	10712090	5.1
FireNet, FireNet+	39999	1.6
SPADE-E2VID	12493030	16.1
HyperE2VID	10149881	6.6
ET-Net	22184698	32.1
Ours	10796853	14.9

[3] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, N. Barnes, L. Kleeman, and R. Mahony, "Reducing the sim-to-real gap for event cameras," in *ECCV*, 2020, pp. 534–549.

[4] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. E. Mahony, and D. Scaramuzza, "Fast image reconstruction with an event camera," in *WACV*, 2020, pp. 156–163.

[5] 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.

[6] 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.

[7] W. Weng, Y. Zhang, and Z. Xiong, "Event-based video reconstruction using transformer," in *ICCV*, 2021, pp. 2543–2552.

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

[9] Y. Jeong, H. Cho, and K.-J. Yoon, "Towards robust event-based networks for nighttime via unpaired day-to-night event translation," in *ECCV*, 2024, pp. 286–306.

[10] Z. Mei, J. Li, B. Zhang, C. Wang, L. Guo, G. Li, and J. Qian, "Temporal-aware spiking transformer hashing based on 3d-dwt," *arXiv preprint arXiv:2501.06786*, 2024.

[11] Y. Fang, Z. Wang, L. Zhang, J. Cao, H. Chen, and R. Xu, "Spiking wavelet transformer," in *ECCV*, 2024, pp. 19–37.

[12] S.-H. Ieng, E. Lehtonen, and R. Benosman, "Complexity analysis of iterative basis transformations applied to event-based signals," *Front. Neurosci*, vol. 12, p. 373, 2018.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[14] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *MICCAI*, 2018, pp. 421–429.

[15] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based optical flow using motion compensation," in *ECCV*, 2019, pp. 711–714.

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

[17] 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 *NeurIPS*, 2015, pp. 802–810.

[18] G. Liang, K. Chen, H. Li, Y. Lu, and L. Wang, "Towards robust event-guided low-light image enhancement: A large-scale real-world event-image dataset and novel approach," in *IEEE/CVF*, 2024, pp. 23–33.

[19] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: an open event camera simulator," in *CoRL*, vol. 87, 2018, pp. 969–982.

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[21] 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.

[22] 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 *CVPR*, 2021, pp. 3445–3454.