HoloEv-Net: Efficient Event-based Action Recognition via Holographic Spatial Embedding and Global Spectral Gating

1st Weidong Hao
Xidian University
Xi'an, China
24251111481@stu.xidian.edu.cn

2nd Rui Fan
Xidian University
Xi'an, China
2017301020002@whu.edu.cn

3rd Juntao Guan*
Xidian University
Xi'an, China
guanjuntao@xidian.edu.cn

4th Rui Lai*
Xidian University
Xi'an, China
rlai@mail.xidian.edu.cn

***Abstract*—Event-based Action Recognition (EAR) has attracted significant attention due to the high temporal resolution and high dynamic range of event cameras. However, existing methods typically suffer from (i) the computational redundancy of dense voxel representations, (ii) structural redundancy inherent in multi-branch architectures, and (iii) the underutilization of spectral information in capturing global motion patterns. To address these challenges, we propose an efficient EAR framework named HoloEv-Net. First, to simultaneously tackle representation and structural redundancies, we introduce a Compact Holographic Spatiotemporal Representation (CHSR). Departing from computationally expensive voxel grids, CHSR implicitly embeds horizontal spatial cues into the Time-Height (T-H) view, effectively preserving 3D spatiotemporal contexts within a 2D representation. Second, to exploit the neglected spectral cues, we design a Global Spectral Gating (GSG) module. By leveraging the Fast Fourier Transform (FFT) for global token mixing in the frequency domain, GSG enhances the representation capability with negligible parameter overhead. Extensive experiments demonstrate the scalability and effectiveness of our framework. Specifically, HoloEv-Net-Base achieves state-of-the-art performance on THU-EACT-50-CHL, HARDVS and DailyDVS-200, outperforming existing methods by 10.29%, 1.71% and 6.25%, respectively. Furthermore, our lightweight variant, HoloEv-Net-Small, delivers highly competitive accuracy while offering extreme efficiency, reducing parameters by 5.4×, FLOPs by 300×, and latency by 2.4× compared to heavy baselines, demonstrating its potential for edge deployment.**

***Index Terms*—event camera, event-based action recognition, FFT**

I. Introduction

Human Action Recognition (HAR) is pivotal for intelligent systems [1]–[3], yet standard frame-based approaches falter under high-speed or High-Dynamic-Range (HDR) conditions due to motion blur and limited dynamic range. Conversely, bio-inspired event cameras represent a paradigm shift. By asynchronously recording pixel-wise brightness changes, they offer microsecond-level latency and superior HDR (>120 dB). These properties make Event-based Action Recognition (EAR) a robust solution for challenging environments where traditional sensors fail.

* Corresponding authors.

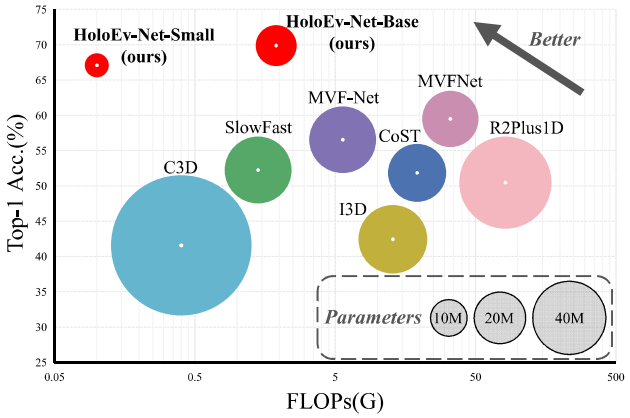

Fig. 1. Performance vs. Cost on DailyDVS-200 [4]. Bubble size denotes model parameters. Our **HoloEv-Net (red)** outperforms existing methods, demonstrating an optimal balance of accuracy and efficiency.

Current state-of-the-art EAR methods [4]–[6] largely rely on voxel representations. While effective, transforming sparse event streams into dense 3D spatiotemporal bins often introduces **computational redundancy** [7]. Processing such high-dimensional inputs necessitates heavy backbones, contradicting the low-latency and energy-efficient nature of neuromorphic sensors. To alleviate this computational burden, multi-view learning, which is a paradigm widely adopted in HAR [8], [9] fuse temporal and spatial dimensions via distinct projections (e.g., Height-Width (H-W), Time-Width (T-W), Time-Height (T-H) views). However, this approach introduces a new bottleneck: **structural redundancy**. Standard multi-view frameworks typically rely on multi-branch architectures, where each view requires a dedicated backbone for feature extraction, leading to significant parameter overhead and computational costs.

To resolve above challenges, we propose the **C**ompact **H**olographic **S**patiotemporal **R**epresentation (**CHSR**). Unlike standard multi-view paradigms that require parallel branches, CHSR achieves multi-view perception within a single view by holographically embedding horizontal spatial cues into the

T-H plane. This strategy yields a compact 2D representation that preserves continuous temporal dynamics and 3D spatial contexts, thereby simultaneously alleviating the computational redundancy of voxel grids and the structural redundancy of multi-branch architectures.

Beyond redundancy, existing frameworks face another limitation: **under-utilization of spectral information**. Human actions inherently possess global periodic patterns [10] that are difficult for lightweight CNNs with limited local receptive fields to capture [11]. While our CHSR effectively transforms these temporal periodicities into distinct wave-like textures within the T-H view (see Fig. 4), standard spatial convolutions are ill-equipped to analyze such global spectral properties.

To address this, we introduce the **G**lobal **S**pectral **G**ating (GSG) module. Capitalizing on the explicit spectral textures exposed by CHSR, GSG leverages the Fast Fourier Transform (FFT) to perform global token mixing in the frequency domain. This mechanism allows the network to efficiently capture long-range dependencies and intrinsic motion periodicities with negligible parameter overhead, perfectly complementing the local operations of the lightweight backbone.

We term this efficient framework **HoloEv-Net**. By integrating the spatially-aware CHSR with the frequency-domain GSG module, our architecture effectively harmonizes representation completeness with model efficiency (Fig. 2). To demonstrate its scalability, we instantiate two variants: a high-performance HoloEv-Net-Base and a lightweight HoloEv-Net-Small. Our main contributions are summarized as follows:

- We propose the Compact Holographic Spatiotemporal Representation (CHSR) to simultaneously dismantle the computational and structural redundancies of existing voxel-based or multi-view paradigms.
- We design a Global Spectral Gating (GSG) module to address CNNs' deficiency in modeling global motion periodicities. Employing FFT in the frequency domain, GSG strengthens feature representation and complements local spatial features with negligible parameter overhead.
- Extensive experiments demonstrate the superiority of our framework. HoloEv-Net-Base achieves SOTA performance on THU-EACT-50-CHL [6], DailyDVS-200 [4], and HARDVS [5]. Meanwhile, HoloEv-Net-Small maintains competitive accuracy while significantly reducing parameters, FLOPs and latency by 5.4×, 300× and 2.4×.

II. Related Work

A. Dense Event Representations

Event cameras generate sparse, asynchronous streams, rendering them incompatible with standard CNNs. Thus, constructing dense representations is a prerequisite for EAR. Early methods [12], [13] aggregated sparse event data along the temporal axis to encode them into 2D images. While bridging the gap to frame-based algorithms, this inevitably compresses rich temporal cues. Furthermore, the voxel grid [14] preserves temporal cues by partitioning the event stream into spatiotemporal bins. This paradigm currently dominates EAR [4]–[6], [15], as dense representations within time windows effectively capture motion evolution. However, processing such high-dimensional structures inherently introduces significant computational redundancy.

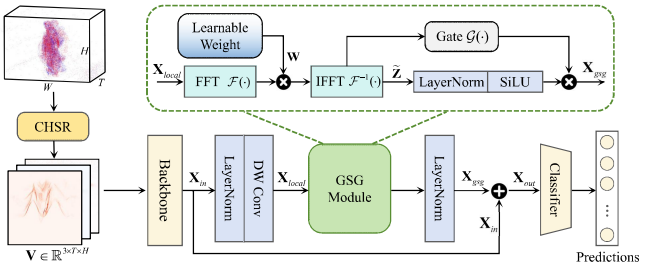

Fig. 2. Overview of the proposed HoloEv-Net. The raw event stream is first processed by our Compact Holographic Spatiotemporal Representation (CHSR). The extracted features from the backbone are processed by the Global Spectral Gating (GSG) module.

B. Multi-View Learning in Action Recognition

To solve this problem, multi-view learning improves action recognition by exploiting complementary spatiotemporal cues from orthogonal projections [8], [9]. However, this paradigm typically relies on multi-branch architectures, where each view requires a dedicated backbone, leading to significant parameter overhead. Furthermore, aggregating these views necessitates complex fusion strategies, ranging from simple concatenation [16] to intricate attention mechanisms [17] or temporal sequence modeling [18]. These explicit fusion designs are often computationally expensive and difficult to optimize. Unlike these heavy architectures, we demonstrate that multi-view perception can be effectively achieved within a single lightweight branch via holographic spatial embedding, eliminating the need for elaborate fusion modules.

C. Frequency Domain Learning in Vision

Frequency analysis is a powerful tool for capturing global dependencies [11], [19]. In Human Activity Recognition (HAR), FTC-Net [19] demonstrated that repetitive activities exhibit distinct spectral signatures, effectively extracted by Fourier Transform Convolutions. For event-based vision, the sensor's microsecond-level temporal resolution makes it ideal for frequency analysis. Recent biological studies [20] have validated that high-frequency periodic behaviors in wildlife can be quantified via Fast Fourier Transform (FFT), proving spectral features often possess greater discriminative power. However, mainstream EAR methods largely overlook global frequency information, relying on local spatial convolutions. Since frequency analysis naturally aligns with our CHSR, we introduce a GSG module to exploit these spectral cues for robust action recognition.

III. Proposed Method

A. Overview

Our proposed **HoloEv-Net**(Fig. 2) processes raw event streams through **C**ompact **H**olographic **S**patiotemporal

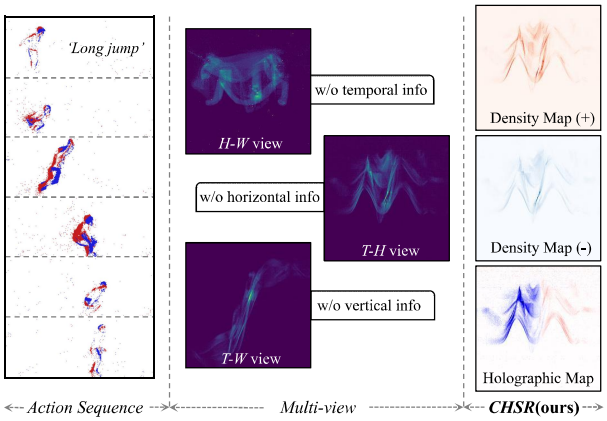

Fig. 3. Visualization of the proposed CHSR. Unlike standard multi-view projections (middle) that suffer from information loss, our CHSR constructs a comprehensive representation in the T-H domain. It consists of three channels: **(1) Density Map (+)** and **(2) Density Map (-)**, which accumulate polarity-specific events to record motion trajectories; and **(3) Holographic Map**, which recovers the horizontal spatial cues (different color denotes different horizontal position) typically lost in T-H projections.

Representation (CHSR), which transforms sparse data into a dense, three-channel 2D tensor while preserving 3D motion contexts. Subsequently, these maps are fed into a lightweight backbone integrated with the **G**lobal **S**pectral **G**ating (**GSG**) module. By leveraging the Fast Fourier Transform (FFT) for global token mixing, GSG captures long-range dependencies often missed by local convolutions. Finally, refined features pass through a classifier for action prediction.

B. CHSR Representation

a) Analysis: Current representation methods for EAR primarily rely on voxel grid, encoding event streams into multiple frames where spatiotemporal cues are embedded between frames. While encoding into H-W, T-W, and T-H views directly embeds spatiotemporal cues within frames, it necessitates complex multi-branch architectures and challenging fusion strategies. To this end, we rethink whether all three views are essential and investigate if complete spatiotemporal cues can be encoded within a single view.

b) Implementation: Among the three views of encoded event streams, the H-W view lacks the temporal cues [14], which is crucial for EAR tasks. Between the temporal views (T-W and T-H), the T-H view is superior as it preserves the vertical topological structure of the human body, directly facilitating action classification. Conversely, the T-W view provides mainly auxiliary information, such as horizontal positioning. As illustrated in Fig. 3 (middle), standard projections inevitably suffer from dimensional information loss (e.g., the T-H view loses horizontal context). Therefore, we embed horizontal information, which comes from T-W view, into the T-H view to form a comprehensive representation. Detailed ablation studies validating this view selection are presented in the next section (Tab. II). Formally, given an input event stream:

$$\mathcal{E} = \{e_k\}_{k=1}^{N} = \{(x_k, y_k, t_k, p_k)\}_{k=1}^{N}, \tag{1}$$

where (x_k, y_k) denotes the spatial coordinates, and t_k represents the timestamp. The polarity $p_k \in \{-1, 1\}$ indicates the sign of brightness change, dividing the stream into positive events ($\mathcal{E}_+$) and negative events ($\mathcal{E}_-$). we construct a dense tensor $\mathbf{V} \in \mathbb{R}^{3\times T\times H}$. The tensor $\mathbf{V}$ is composed of three complementary feature maps: positive density map ($\mathcal{D}^+$), negative density map ($\mathcal{D}^-$) and holographic map ($\mathcal{H}$), as visualized in the right column of Fig. 3.

$$\mathcal{D}^{\pm}(t, y) \doteq \sum_{e_k \in \mathcal{E}_{\pm}} \delta(t - t_k, y - y_k), \tag{2}$$

where δ represents the Kronecker delta function. By aggregating the impulses based on event polarity, this function effectively accumulates the local event density over time in the T-H view.

$$\mathcal{H}(t, y) \doteq \sum_{k=1}^{N} \phi(x_k) \cdot \delta(t - t_k, y - y_k), \tag{3}$$

where $\phi(x_k) = sin(\pi x_k / W)$ denotes the transverse spatial embedding function. By transforming horizontal coordinates into cyclic sinusoidal patterns, this operation implicitly preserves essential 3D spatiotemporal contexts within the 2D T-H view, effectively realizing a holographic encoding mechanism.

C. GSG Module

a) Analysis: As illustrated in Fig. 4, human actions exhibit distinct periodic oscillatory patterns in the T-H view, implying high discriminability in the frequency domain. While standard CNNs excel at extracting local textures, their limited receptive fields struggle to capture these global periodic dependencies. To address this, we propose the Global Spectral Gating (GSG) module, which performs efficient global token mixing in the frequency domain.

b) Architecture Overview: As shown in Fig. 2, our framework is instantiated with either a ResNet-18 [26] or a MobileNetV3-Small [27] backbone to process the CHSR tensor $\mathbf{V} \in \mathbb{R}^{3\times T\times H}$. The backbone extracts high-level features, denoted as $\mathbf{X}_{in} \in \mathbb{R}^{C\times T'\times H'}$. We integrate the GSG module at the end of the backbone via a global residual connection. This design fuses the local spatial features from the backbone with the spectrally modulated global context, formulated as $\mathbf{X}_{out} = \mathbf{X}_{in} + \mathbf{X}_{gsg}$. This residual structure ensures robust feature learning and stable gradient flow.

c) Implementation: The GSG module operates through three sequential stages: **(1) Local Embedding.** We first apply a 3×3 Depthwise Convolution (DWConv) to $\mathbf{X}_{in}$. This step encodes local spatial contexts and prevents high-frequency spectral aliasing.**(2) Global Spectral Filtering.** To capture long-range dependencies, the features are transformed into the frequency domain via a 2D Real-to-Complex FFT along the temporal (T) and height (H) axes. A learnable complex weight

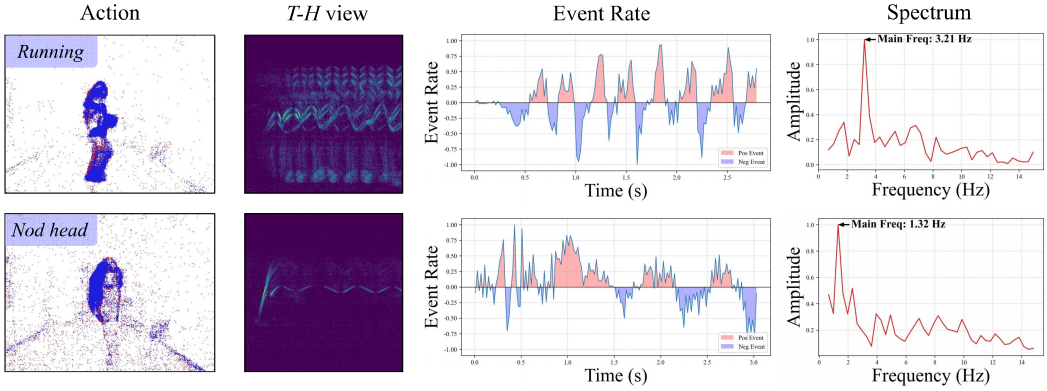

Fig. 4. Frequency analysis of event streams. The figure displays the event rate and FFT spectrum for two distinct actions. The distinct main frequencies (3.21 Hz for running, 1.32 Hz for nodding) highlight the discriminability of actions in the frequency domain.

TABLE I
COMPARISONS ON THU-EACT-50-CHL, HARDVS AND DAILYDVS-200. BEST RESULTS IN **BOLD** AND SECOND BEST UNDERLINED.

Methods	Input Type	Backbone	THU-EACT-50-CHL		HARDVS		DailyDVS-200		FLOPs	Params	Latency
			Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)			
R2Plus1D [21]	Frame	ResNet34	50.54	73.15	49.06	56.43	36.06	63.67	81.4G	63.5M	11.2ms
SlowFast [22]		ResNet50	46.14	73.16	50.63	57.77	41.49	68.19	1.4G	33.6M	19.6ms
TSM [23]		ResNet50	44.30	71.31	52.63	60.56	40.87	71.46	1.4G	24.3M	11.8ms
I3D [1]		ResNet50	42.41	68.49	51.53	59.82	32.30	59.05	12.8G	35.3M	13.1ms
C3D [24]		3D CNN	41.54	68.57	50.52	56.14	21.99	45.81	0.4G	147.2M	8.4ms
EST [25]	Learnable	ResNet34	55.07	76.02	36.51	42.09	32.23	59.66	4.2G	21.5M	11.7ms
CoST [8]	Multi-view	ResNet50	51.84	75.00	50.57	61.38	36.09	64.45	19.2G	25.4M	21.3ms
MVFNet [9]		ResNet50	59.56	77.49	52.98	63.26	48.30	75.91	32.8G	23.6M	17.5ms
MVF-Net [16]		ResNet34&18	56.47	76.53	52.61	61.67	43.98	70.39	5.6G	33.6M	13.7ms
HoloEv-Net-Base	Holographic	ResNet-18	**69.85**	**79.60**	**54.69**	**63.56**	**54.55**	**77.50**	1.9G	12.5M	9.6ms
HoloEv-Net-Small		MobileNetV3-Small	67.10	78.86	52.69	62.56	49.13	73.56	**0.1G**	**4.4M**	**7.4ms**

tensor $\mathbf{W}$ modulates the spectrum to emphasize action-specific frequencies:

$$\tilde{\mathbf{Z}} = \mathcal{F}^{-1}\left(\mathcal{F}(\mathbf{X}_{local}) \odot \mathbf{W}\right), \tag{4}$$

where $\mathcal{F}(\cdot)/\mathcal{F}^{-1}(\cdot)$ denote FFT/Inverse FFT. This operation achieves global information interaction.**(3) Gated Reconstruction.** To suppress spectral artifacts, the refined features $\tilde{\mathbf{Z}}$ pass through a dual-branch gating mechanism. One branch acts as a content carrier (normalized by LayerNorm and activated by SiLU), while the other serves as a dynamic gate $\mathcal{G}(\cdot)$:

$$\mathbf{X}_{gsg} = \text{SiLU}(\text{LN}(\tilde{\mathbf{Z}})) \odot \mathcal{G}(\tilde{\mathbf{Z}}). \tag{5}$$

This gating mechanism selectively filters noise while retaining salient motion features before the final residual fusion.

IV. EXPERIMENT

A. Implementation Details

a) Datasets: To evaluate the effectiveness and robustness of our proposed method, we conduct comprehensive experiments on three challenging event-based action recognition benchmarks: THU-EACT-50-CHL [6], HARDVS [5] and DailyDVS-200 [4]. THU-EACT-50-CHL contains 2,330 samples across 50 categories, featuring distinct lighting conditions. HARDVS is a large-scale dataset comprising 107,646 sequences of 300 action classes performed by 5 subjects. DailyDVS-200 includes 22,046 sequences covering 200 daily action categories, characterized by 14 environmental challenges such as camera motion and illumination changes.

b) Training Protocol: All experiments were implemented using the PyTorch framework and conducted on a single NVIDIA RTX 3090 GPU. Following [16], the T axis is discretized into T = 224. The network was trained for 50 epochs using the Adam optimizer with an initial learning rate of $1e-4$. The batch size is set to 32. We used the Cross-Entropy loss function for optimization.

c) Evaluation Metrics: To measure recognition accuracy, we report Top-1 and Top-5 accuracy. Beyond accuracy, to assess the model's complexity and computational efficiency, we provide model parameters (Params) and floating-point operations (FLOPs). Furthermore, we measure the inference Latency (ms/sample) to evaluate suitability for edge deployment.

TABLE II
ABLATION STUDY ON VIEW SELECTION AND EMBEDDING STRATEGIES ON THU-EACT-50-CHL.

Group	View Combination	Top-1 (%)	FLOPs (G)
Single-view	H-W	38.60	1.82
	T-W	50.18	1.82
	T-H	60.36	1.82
Multi-view	H-W + T-H	63.91	1.90
	H-W + T-W	54.23	1.90
	T-W + T-H	65.07	1.90
	H-W + T-H + T-W	<u>65.09</u>	1.98
Holographic	T-W with H	60.16	1.86
	T-H with W	**69.85**	1.86

TABLE III
ABLATION STUDY ON THE GSG COMPONENTS ON THU-EACT-50-CHL.

Component	Variants			
Baseline (ResNet)	✓	✓	✓	✓
FFT		✓	✓	✓
Learnable Weight			✓	✓
Gated Reconstruction				✓
Top-1 Acc (%)	65.07	66.36	<u>68.19</u>	**69.85**

TABLE IV
ANALYSIS AGAINST CAMERA MOTION ON DAILYDVS-200.

Method	Camera Movement	
	Static (%)	Motion (%)
SlowFast [22]	51.81	21.05
TSM [23]	50.16	26.88
EST [25]	41.51	13.52
Spikeformer [28]	46.15	13.74
Swin-T [29]	<u>58.05</u>	<u>27.84</u>
HoloEv-Net-Base (w/o GSG)	**58.59**	**29.31**
HoloEv-Net-Base	**59.62**	**38.26**

TABLE V
ABLATION STUDY ON THE INTEGRATION STAGE OF THE GSG MODULE WITHIN THE RESNET-18 BACKBONE.

Insertion Stage	Feature Resolution	Top-1 Acc (%)
Baseline (None)	-	65.07
Stage 1	56×56	66.54
Stage 2	28×28	66.72
Stage 3	14×14	<u>67.83</u>
Stage 4 (Ours)	$\mathbf{7 \times 7}$	**69.85**

B. Quantitative Results

We compare our HoloEv-Net with recent state-of-the-art approaches, including frame-based [1], [21]–[24], learnable [25], and multi-view [8], [9], [16] methods. The quantitative results on THU-EACT-50-CHL, HARDVS and DailyDVS-200 are summarized in Tab. I.

a) Performance Improvement: Our HoloEv-Net-Base achieves 69.85%, 54.69% and 54.55% Top-1 accuracy on three datasets, outperforming the leading method MVFNet [9] by 10.29%, 1.71% and 6.25%, respectively.

b) Efficiency Analysis: In terms of computational cost, our HoloEv-Net-Base requires only 12.5M parameters and 1.86G FLOPs (measured on DailyDVS-200). Furthermore, our HoloEv-Net-Small pushes the limits of efficiency with only 4.4M parameters and 0.1G FLOPs. Crucially, as shown in Table I, HoloEv-Net-Small achieves a remarkable latency of 7.4 ms on RTX 3090. More importantly, on the embedded NVIDIA Jetson AGX Orin, it sustains a real-time latency of 32.4 ms, confirming its suitability for edge deployment.

C. Ablation Study

a) Analysis on Representation Construction: We evaluate projection views and encoding strategies by comparing single-view, multi-view fusion, and our holographic representation (Tab. II). For fair comparison, backbone input layers are adjusted accordingly. Among single-view, H-W yields the lowest accuracy (38.60%), confirming that lacking temporal dimensions critically harms EAR. Furthermore, T-H consistently outperforms$T$-W (+10.18%). Standard multi-view methods (similar to MVF-Net [16]) attempt to combine information, they suffer from dimensional misalignment—directly concatenating features from heterogeneous domains (e.g. spatial H vs. temporal T) introduces optimization difficulties and limits performance gains. Consequently, our Holographic encoding (T-H with W) proves optimal. Embedding horizontal motion cues into the T-H view efficiently couples spatiotemporal information while circumventing traditional dimensional mismatches.

b) Analysis on GSG: Evaluating components (Tab. III), FFT improves the baseline by 1.29% (to 66.36%), learnable weights add 1.83%, and gated reconstruction achieves 69.85% Top-1 accuracy. Furthermore, GSG ensures stability in dynamic environments. Under significant camera motion (Tab. IV), GSG-equipped HoloEv-Net-Base outperforms Swin-T and its GSG-free counterpart by 10.42% and 8.95% respectively, proving its efficacy in filtering motion-induced noise while preserving salient features.

c) Analysis on Module Integration Stages: Finally, we investigate the optimal position for the GSG module by inserting it after each stage of the ResNet-18 backbone (Tab. V). The feature resolution indicates the feature map size at each stage. We observe that the performance improvement becomes more significant as the module is inserted into deeper layers, peaking at Stage 4 (69.85%). Early-stage global filtering risks blurring high-resolution visual primitives or amplifying high-frequency noise. In contrast, deep-stage integration leverages global receptive fields to capture crucial long-range dependencies and temporal rhythms. Moreover, processing lower-resolution feature maps at Stage 4 (7×7) maximizes computational efficiency.

V. CONCLUSION

In this paper, we proposed efficient HoloEv-Net for EAR task. To address computational and structural redundancy, we

introduced CHSR. This strategy implicitly achieves multi-view perception within a single T-H view by holographically embedding horizontal spatial cues into the channel dimension. Furthermore, to utilize spectral information of EAR, we designed a Global Spectral Gating (GSG) module, which leverages the FFT to capture long-range dependencies with negligible parameter overhead. Extensive experiments on three benchmark datasets demonstrate the superiority of our framework. HoloEv-Net-Base achieves SOTA accuracy. Meanwhile, HoloEv-Net-Small, pushes the limits of efficiency, achieving low latency suitable for real-time edge deployment.

While achieving SOTA, the modest gains on HARDVS highlight the limitations of our spectral gating in complex scenarios. Future work will focus on adaptive frequency learning to dynamically accommodate diverse motion patterns. Furthermore, we aim to generalize the holographic paradigm to downstream tasks such as object detection and optical flow.

VI. Acknowledgements

This work was supported in part by the Key Project of National Natural Science Foundation of China (No.62534006), in part by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No.JYB2025XDXM121), in part by the Young Scientists Fund of the National Natural Science Foundation of China (No.62304162), in part by the China Postdoctoral Science Foundation (No.2024M762532), and in part by the Shaanxi Provincial Natural Science Foundation for Basic Research Program (No.2024JC-YBMS-794).

References

[1] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[2] D. P. Losey, K. Srinivasan, A. Mandlekar, A. Garg, and D. Sadigh, "Controlling assistive robots with learned latent actions," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 378–384.

[3] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European conference on computer vision*. Springer, 2016, pp. 510–526.

[4] Q. Wang, Z. Xu, Y. Lin, J. Ye, H. Li, G. Zhu, S. A. Ali Shah, M. Bennamoun, and L. Zhang, "Dailydvs-200: A comprehensive benchmark dataset for event-based action recognition," in *European Conference on Computer Vision*. Springer, 2024, pp. 55–72.

[5] X. Wang, Z. Wu, B. Jiang, Z. Bao, L. Zhu, G. Li, Y. Wang, and Y. Tian, "Hardvs: Revisiting human activity recognition with dynamic vision sensors," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 6, 2024, pp. 5615–5623.

[6] Y. Gao, J. Lu, S. Li, N. Ma, S. Du, Y. Li, and Q. Dai, "Action recognition and benchmark using event cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14 081–14 097, 2023.

[7] R. Fan, W. Hao, J. Guan, L. Rui, L. Gu, T. Wu, F. Zeng, and Z. Zhu, "Eventpillars: Pillar-based efficient representations for event data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 3, 2025, pp. 2861–2869.

[8] C. Li, Q. Zhong, D. Xie, and S. Pu, "Collaborative spatiotemporal feature learning for video action recognition," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2019, pp. 7872–7881.

[9] W. Wu, D. He, T. Lin, F. Li, C. Gan, and E. Ding, "Mvfnet: Multi-view fusion network for efficient video recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 2943–2951.

[10] T. F. Runia, C. G. Snoek, and A. W. Smeulders, "Real-world repetition estimation by div, grad and curl," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9009–9017.

[11] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," *Advances in neural information processing systems*, vol. 34, pp. 980–993, 2021.

[12] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5419–5427.

[13] F. Cladera, A. Bisulco, D. Kepple, V. Isler, and D. D. Lee, "On-device event filtering with binary neural networks for pedestrian detection using neuromorphic vision sensors," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3084–3088.

[14] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Ev-flownet: Self-supervised optical flow estimation for event-based cameras," *arXiv preprint arXiv:1802.06898*, 2018.

[15] X. Wang, S. Wang, P. Shao, B. Jiang, L. Zhu, and Y. Tian, "Event stream based human action recognition: a high-definition benchmark dataset and algorithms," *arXiv preprint arXiv:2408.09764*, 2024.

[16] Y. Deng, H. Chen, and Y. Li, "Mvf-net: A multi-view fusion network for event-based object classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8275–8284, 2021.

[17] N. Song, S. Du, Z. Wu, L. Zhong, L. T. Yang, J. Yang, and S. Wang, "Gaf-net: Graph attention fusion network for multi-view semi-supervised classification," *Expert Systems with Applications*, vol. 238, p. 122151, 2024.

[18] A. Ullah, K. Muhammad, T. Hussain, and S. W. Baik, "Conflux lstms network: A novel approach for multi-view action recognition," *Neurocomputing*, vol. 435, pp. 321–329, 2021.

[19] J. Xie, Y. Wu, W. Li, M. Yang, R. Hou, and Y. Li, "Ftc-net: Fourier transform convolution for multivariate time series classification in human activity recognition," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8.

[20] F. Hamann, S. Ghosh, I. Juárez Martínez, T. Hart, A. Kacelnik, and G. Gallego, "Fourier-based action recognition for wildlife behavior quantification with event cameras," *Advanced Intelligent Systems*, vol. 7, no. 2, p. 2400353, 2025.

[21] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[22] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[23] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7083–7093.

[24] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[25] D. Gehrig, A. Loquercio, K. G. Derpanis, and D. Scaramuzza, "End-to-end learning of representations for asynchronous event-based data," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5633–5643.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[28] Z. Zhou, Y. Zhu, C. He, Y. Wang, S. Yan, Y. Tian, and L. Yuan, "Spikformer: When spiking neural network meets transformer," *arXiv preprint arXiv:2209.15425*, 2022.

[29] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.