RIFE-AOV: An Efficient Real-time Video Frame Interpolation Method for Always-on Video

Xin Chen[1], Daqing Chen[1], Jun Lei[1*], Liquan Shen[2], Nan Wu[1]
[1] *China Mobile (Hangzhou) Information Technology Co., Ltd. Hangzhou, China*
[2] *Shanghai University, Shanghai, China*
Email: chenxinhz@cmhi.chinamobile.com, chendaqing@cmhi.chinamobile.com,
leijun@cmhi.chinamobile.com, jsslq@shu.edu.cn, wunan@cmhi.chinamobile.com

***Abstract*—In Always-on Video (AOV) systems, switching from a low frame-rate (LFR) mode to a high frame-rate (HFR) mode often leads to temporal discontinuities in video sequences. Existing video frame interpolation (VFI) methods are not specifically designed for this scenario, and therefore tend to suffer from information loss and blurred motion boundaries during frame-rate transitions. To address these issues, this paper proposes RIFE-AOV, an efficient real-time frame interpolation method tailored for AOV applications. Built upon the RIFE framework, RIFE-AOV targets the limitations of the optical flow estimation network IFNet under frame-rate switching conditions by structurally enhancing its core sub-module, IFBlock, with a Global Directional Attention Mechanism (GDAM). Specifically, Global Channel Attention (GCA) is employed to emphasize critical motion features and alleviate information loss, while Directional Spatial Attention (DSA) strengthens the modeling of motion boundaries and structural directionality, thereby reducing motion boundary blurring. Experiments conducted on the Vimeo90K dataset demonstrate that RIFE-AOV achieves superior interpolation performance in dynamic scenes, in terms of PSNR and SSIM, compared with state-of-the-art real-time VFI methods, while maintaining real-time efficiency. These results validate the practical effectiveness of the proposed method for AOV systems.**

***Index Terms*—Video Frame Interpolation, Always-on Video, Optical Flow, Attention Mechanism, Real-Time Video Processing**

I. Introduction

With the rapid development of the Internet of Things (IoT) and intelligent surveillance technologies, network cameras (IP cameras) have been widely deployed in public security, intelligent transportation, smart cities, and smart home applications [1]. However, the continuous generation of high frame-rate video streams from massive numbers of devices places substantial pressure on network bandwidth and cloud storage resources, resulting in high operational costs. Consequently, low-power and high-efficiency intelligent video technologies have become a major focus of industrial research, among which Always-on Video (AOV) has emerged as a practical solution.

AOV allows cameras to operate in an ultra-low frame-rate standby mode (e.g., 5 fps), while switching to full frame-rate recording (e.g., 15 fps) only when specific targets, such as humans or vehicles, are detected. This event-driven mechanism effectively reduces redundant data transmission and storage in predominantly static scenes, making AOV particularly suitable for large-scale and long-term surveillance scenarios.

Despite its advantages, AOV introduces a critical challenge during the transition from low frame-rate (LFR) standby mode to high frame-rate (HFR) recording mode. When an event is triggered at time T, the camera cannot immediately generate temporally dense frames, and the initial period after activation still relies on sparsely sampled inputs. This abrupt frame-rate discontinuity leads to noticeable visual stuttering and motion inconsistency during playback, potentially obscuring important motion details and compromising evidential reliability.

Video Frame Interpolation (VFI), which synthesizes intermediate frames between adjacent inputs to improve temporal continuity, offers a feasible solution for alleviating low frame-rate artifacts in AOV scenarios. Recently, deep learning-based real-time VFI methods have made notable progress in balancing visual quality and computational efficiency. Among them, Real-time Intermediate Flow Estimation (RIFE) directly estimates intermediate optical flow and employs mask-based fusion, achieving a favorable trade-off between accuracy and speed [2]. However, under AOV low frame-rate triggering conditions, RIFE still faces two major limitations: **(i)** significant **information loss** caused by large temporal gaps between input frames, increasing motion uncertainty and leading to missing details; and **(ii)** noticeable **motion boundary blurring**, where fast-moving object edges are over-smoothed, resulting in structural degradation.

To address these challenges, we propose **RIFE-AOV**, an enhanced real-time video frame interpolation method tailored for AOV scenarios. Effective interpolation under frame-rate switching requires strengthened motion representation under large temporal gaps, accurate preservation of motion boundaries, and real-time efficiency on resource-constrained devices. Guided by this insight, RIFE-AOV improves the feature modeling capability of the optical flow estimation module IFNet within the RIFE framework through lightweight yet effective architectural adaptations.

The main contributions of this work are summarized as follows:

- We propose **RIFE-AOV**, an enhanced real-time video frame interpolation method tailored for Always-on Video

This work was supported by the National Natural Science Foundation of China (Nos. NSFC U22B2001)
* Corresponding author.

applications. By integrating AOV-aware feature modeling strategies into the core IFBlock of the optical flow estimation module IFNet within the RIFE framework, the proposed method significantly improves visual continuity immediately after event triggering.

- We introduce a lightweight **Global Channel Attention (GCA)** mechanism. Based on Global Average Pooling (GAP), GCA captures channel-wise motion statistics and performs adaptive channel reweighting, strengthening motion-relevant representations and mitigating intermediate frame detail loss.
- We propose a **Directional Spatial Attention (DSA)** mechanism to explicitly model motion boundary orientation using horizontally and vertically decomposed convolutions. This design enhances sensitivity to regions with strong spatial gradients, effectively reducing motion boundary blurring and structural artifacts in interpolated frames.

II. Related Work

Video frame interpolation (VFI) aims to synthesize intermediate frames between adjacent video frames to improve temporal continuity [3]. Existing methods can be broadly categorized into traditional model-based approaches and deep learning-based approaches [4]. In recent years, real-time interpolation methods have also attracted increasing attention due to their practical relevance in resource-constrained scenarios. This section reviews representative work in these categories and discusses their limitations in Always-on Video low frame-rate triggering scenarios.

A. Traditional Video Frame Interpolation Methods

Traditional VFI methods typically follow a pipeline of motion estimation and motion compensation. Early approaches estimate pixel-wise optical flow between consecutive frames and warp the input frames to synthesize intermediate results [5], [6]. Various extensions have been proposed to improve robustness, including bidirectional flow estimation, occlusion reasoning, confidence-aware blending, and regularization strategies [7].

Despite their clear physical interpretability and relatively low computational complexity, traditional methods are highly sensitive to optical flow accuracy. In the presence of large motion displacement, complex occlusions, or illumination variations, flow estimation errors often lead to ghosting artifacts and structural distortions. Moreover, their reliance on small temporal gaps makes them ill-suited for handling the sparse and discontinuous frame sampling commonly encountered in AOV low frame-rate inputs.

B. Deep Learning-Based Video Frame Interpolation

With the advancement of deep learning, data-driven VFI methods have substantially improved interpolation quality by learning complex motion patterns directly from data. Representative approaches such as Super SloMo [8], DAIN [9], and SepConv [10] employ convolutional neural networks to jointly estimate motion and synthesize intermediate frames. Subsequent works further enhance performance by incorporating adaptive kernels, depth cues, context-aware synthesis, or deformable convolutions [11], [12]. More recently, transformer-based and flow-free methods, including CAIN [13], VFIformer [15], and AMT [16], have been proposed to capture long-range dependencies and implicit motion representations. In addition, large-scale pretraining and multi-frame interpolation strategies have been explored to further improve reconstruction fidelity [17].

To balance interpolation quality and efficiency, a series of real-time deep learning-based VFI methods have been proposed. Representative approaches such as Real-time Intermediate Flow Estimation (RIFE) [2] directly predict intermediate optical flow and employ mask-based blending to synthesize intermediate frames, avoiding costly multi-stage refinement. Lightweight variants, including IFRNet [14] and other efficient architectures [18], [19], further reduce latency through architectural simplification. Recent studies have also explored improving temporal consistency and robustness under real-time constraints, for example by refining intermediate features or introducing efficient attention mechanisms [20].

While these deep learning-based methods achieve impressive performance on standard benchmarks with dense high frame-rate inputs, their effectiveness degrades noticeably under sparse sampling with large temporal gaps, as commonly observed in Always-on Video scenarios. Furthermore, many methods rely on heavy architectures and high computational overhead, limiting their applicability in real-time or edge deployments.

C. Limitations in AOV Low Frame-Rate Scenarios

AOV systems introduce a unique interpolation challenge characterized by abrupt transitions from ultra-low frame-rate standby mode to high frame-rate recording mode. During the initial period after event triggering, available frames remain sparse and discontinuous, violating the assumptions of most existing VFI methods. Low frame-rate inputs exacerbate motion estimation errors, while fast-moving objects are prone to boundary over-smoothing and structural degradation. Although recent studies have explored robustness under large motion or occlusion, VFI methods explicitly tailored for AOV low frame-rate triggering scenarios remain largely unexplored.

Motivated by these observations, this work focuses on enhancing the feature modeling capability of real-time optical flow-based interpolation frameworks to better accommodate AOV-specific constraints. By introducing lightweight attention mechanisms into the optical flow estimation module, the proposed method aims to mitigate information loss and motion boundary blurring while preserving real-time performance.

III. Method

RIFE is a real-time video frame interpolation framework that synthesizes intermediate frames by estimating intermediate motion and adaptively blending information from two consecutive input frames. Its optical flow estimation network,

IFNet, adopts a cascaded refinement strategy to directly predict intermediate bidirectional optical flows together with a blending mask, progressively refining motion representations in a coarse-to-fine manner without explicitly estimating forward and backward flows between input frames. The fundamental building unit of IFNet is the IFBlock, a lightweight convolutional refinement module, and multiple IFBlocks are stacked to iteratively refine optical flow and mask predictions. In this work, we focus on enhancing IFBlock to better handle challenging motion patterns under large temporal gaps, while preserving the overall efficiency of the RIFE framework. An overview of the architecture and the proposed IFBlock enhancement is presented in the following subsections.

A. RIFE Frame Interpolation Framework

RIFE performs end-to-end video frame interpolation without explicit motion trajectory modeling. Given two consecutive input frames $\mathbf{I}_0, \mathbf{I}_1 \in \mathbb{R}^{H\times W\times 3}$, IFNet predicts intermediate bidirectional optical flows $\mathbf{F}_{0\to t}, \mathbf{F}_{1\to t} \in \mathbb{R}^{H\times W\times 2}$ and a blending mask $\mathbf{M} \in [0,1]^{H\times W}$. The intermediate frame $\hat{\mathbf{I}}_t$ is synthesized as

$$\hat{\mathbf{I}}_t = \mathbf{M} \odot \mathcal{W}(\mathbf{I}_0, \mathbf{F}_{0\to t}) + (1-\mathbf{M}) \odot \mathcal{W}(\mathbf{I}_1, \mathbf{F}_{1\to t}), \quad (1)$$

where $\mathcal{W}(\cdot)$ denotes the bilinear warping operator and $\odot$ represents element-wise multiplication.

Within IFNet, the IFBlock serves as the core computational unit. Each IFBlock takes frame features, current optical flow estimates, and mask predictions as inputs, and outputs residual updates for optical flow $\Delta\mathbf{F}$ and blending mask $\Delta\mathbf{M}$ via convolutional feature extraction and residual learning.

B. Overall Design of IFBlock Enhancement

Although the original IFBlock in RIFE is carefully optimized for real-time inference, its lightweight convolutional design mainly relies on local feature aggregation. This limitation becomes evident in AOV scenarios, where large temporal gaps between sparsely sampled frames introduce large motion displacement, occlusion, and ambiguous motion boundaries. Consequently, interpolated frames often suffer from missing fine details and blurred object edges, particularly in regions with fast or complex motion.

To improve robustness under these conditions without significantly increasing computational overhead, we enhance the feature representation capability of IFBlock by introducing a lightweight attention-based module that explicitly incorporates global contextual modeling and directional spatial awareness. This module enables IFBlock to selectively emphasize motion-relevant channels while capturing spatial dependencies aligned with dominant motion structures, thereby suppressing background noise and preserving sharp motion boundaries. We term this plug-in enhancement as the Global and Directional Attention Module (GDAM), which is integrated into IFBlock while preserving the original cascaded refinement architecture of IFNet and improving motion modeling capability under AOV frame-rate switching scenarios.

The enhanced IFBlock inserts the proposed GDAM module after the original downsampling convolution layers, as illustrated in Fig. 1.

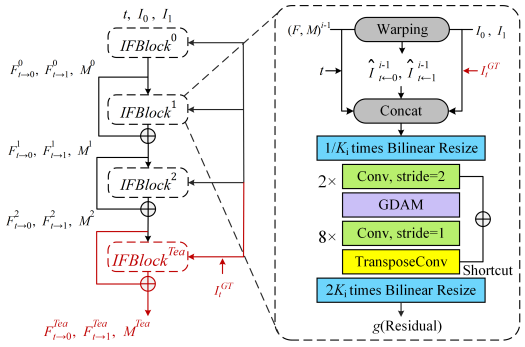

Fig. 1. The structure of IFBlock within IFNet. **Left:** IFNet stacks multiple IFBlocks operating at different resolutions. **Right:** In each IFBlock, input frames are backward-warped using the current flow estimate, then combined with previous-stage outputs and the interpolation timestep to refine residual flow and blending mask.

Inspired by the Global Attention Mechanism (GAM) [21], GDAM is composed of two serial submodules: Global Channel Attention (GCA) and Directional Spatial Attention (DSA), which jointly enhance channel-wise discriminative information and spatial directional awareness, as shown in Fig. 2.

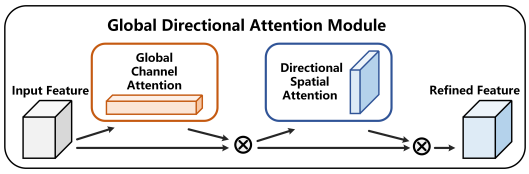

Fig. 2. The structure of GDAM.

C. Global Channel Attention (GCA)

1) Information Loss Analysis: In optical flow estimation, different feature channels encode distinct visual cues such as edges, textures, and motion patterns. Conventional convolution operations aggregate channel information as

$$\mathbf{Y} = \sum_{c=1}^{C} \mathbf{W}_c * \mathbf{X}_c, \quad (2)$$

where $\mathbf{X}_c$ denotes the c-th channel of the input feature map and $\mathbf{W}_c$ represents the corresponding convolution kernel. This implicit channel averaging ignores inter-channel importance and may suppress motion-relevant features, resulting in missing fine details in interpolated frames.

2) GCA Modeling: In the cascaded refinement process of IFBlock, the extracted feature channels encode diverse information, including motion cues, appearance details, and warped residuals. Under large temporal gaps and complex motion patterns, which are common in AOV scenarios, these channels contribute unequally to accurate intermediate flow estimation. Uniformly treating all channels may introduce

redundant or noisy responses, thereby limiting detail preservation and motion boundary sharpness in the interpolated frames.

To address this limitation, we incorporate a GCA that adaptively emphasizes informative feature channels while suppressing less relevant ones. Such a design enables IFBlock to focus on motion and structure sensitive representations, as illustrated in Fig. 3.

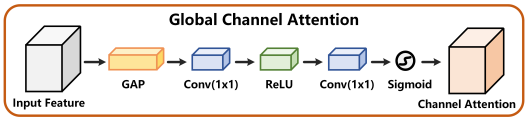

Fig. 3. The structure of GCA.

Given an input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, Global Average Pooling (GAP) aggregates spatial information into a channel descriptor:

$$z_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{X}_c(i,j), \quad c = 1,\ldots,C. \tag{3}$$

Two successive 1×1 convolutions are then applied to capture nonlinear inter-channel dependencies:

$$\mathbf{s} = \sigma\left(\mathbf{W}_2\,\delta\left(\mathbf{W}_1\mathbf{z}\right)\right), \tag{4}$$

where $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ are learnable parameters, r is the channel reduction ratio, $\delta(\cdot)$ denotes the ReLU activation, and $\sigma(\cdot)$ denotes the Sigmoid function. The channel-enhanced features are obtained via

$$\mathbf{X}_c^{\mathrm{CA}} = s_c\cdot\mathbf{X}_c. \tag{5}$$

3) Analysis: GCA adaptively estimates the contribution of each channel to motion representation, suppressing redundant channels while emphasizing motion- and edge-sensitive features. Compared to fully connected channel modeling, the GAP plus 1×1 convolution design introduces minimal parameters and computational overhead, making it suitable for real-time interpolation.

D. Directional Spatial Attention (DSA)

1) Motion Boundary Blur Analysis: Conventional spatial attention mechanisms, such as those using isotropic 7×7 convolutions, tend to smooth high-frequency gradients:

$$\nabla(\mathbf{K}_{7\times 7} * \mathbf{X}) \approx \mathbf{K}_{7\times 7} * \nabla\mathbf{X}, \tag{6}$$

which makes it difficult to preserve sharp motion boundaries. However, motion boundaries are inherently directional, and isotropic spatial attention often leads to boundary blurring and ghosting artifacts.

2) DSA Modeling: In video frame interpolation, spatial attention is critical for preserving motion boundaries and structural continuity in the synthesized frames. In AOV scenarios with large temporal gaps, motion boundaries often exhibit strong directional characteristics due to object displacement and occlusion, making them difficult to model using conventional isotropic spatial attention mechanisms. Treating all spatial directions uniformly may blur motion edges and weaken structural coherence in regions with complex or fast motion.

To explicitly enhance directional sensitivity, we introduce DSA to capture orientation-aware spatial dependencies within feature maps. By modeling horizontal and vertical directional responses separately, DSA enables IFBlock to emphasize spatial locations that are highly correlated with motion edges and structural transitions, as illustrated in Fig. 4.

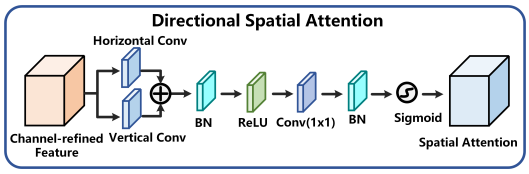

Fig. 4. The structure of DSA.

Given channel-enhanced features $\mathbf{X}^{\mathrm{CA}}$, directional convolutions are applied:

$$\mathbf{H} = \mathbf{W}_h * \mathbf{X}^{\mathrm{CA}}, \quad \mathbf{V} = \mathbf{W}_v * \mathbf{X}^{\mathrm{CA}}, \tag{7}$$

where $\mathbf{W}_h$ is a 1×3 convolution kernel approximating the horizontal gradient $\partial/\partial x$, and $\mathbf{W}_v$ is a 3×1 kernel approximating the vertical gradient $\partial/\partial y$. The directional features are fused as

$$\mathbf{S} = \mathbf{H} + \mathbf{V}. \tag{8}$$

A 1×1 convolution followed by a Sigmoid activation generates the spatial attention map:

$$\mathbf{A}_s = \sigma\left(f_{1\times 1}\left(\delta(\mathbf{S})\right)\right). \tag{9}$$

The final spatially weighted features are obtained by

$$\mathbf{X}^{\mathrm{SA}} = \mathbf{X}^{\mathrm{CA}} \odot \mathbf{A}_s. \tag{10}$$

3) Analysis: From image gradient theory, motion boundaries correspond to extrema of pixel intensity gradients:

$$\|\nabla I\| = \sqrt{\left(\frac{\partial I}{\partial x}\right)^2 + \left(\frac{\partial I}{\partial y}\right)^2}. \tag{11}$$

By explicitly enhancing horizontal and vertical gradient responses, DSA assigns higher attention weights to motion boundary regions, reducing cross-boundary feature mixing and effectively alleviating motion boundary blur in interpolated frames.

E. Summary

By integrating Global Channel Attention and Directional Spatial Attention into IFBlock, the proposed RIFE-AOV method effectively mitigates detail loss and motion boundary blur in complex motion scenarios while maintaining real-time efficiency. The enhanced IFBlock improves optical flow estimation accuracy and leads to higher-quality intermediate frame synthesis without introducing significant computational overhead.

IV. Experiments

A. Experimental Settings

All experiments were conducted on a server equipped with four NVIDIA V100 GPUs. Model training and evaluation were performed on the widely used Vimeo90K dataset [22], which is specifically designed for video frame interpolation and related video restoration tasks. Vimeo90K consists of high-quality video clips collected from online videos and is organized into short sequences containing consecutive frames. Following the standard protocol, the dataset is divided into a training set and a test set, where the training set contains 51,312 triplets for model optimization, and the test set includes 3,782 triplets for quantitative evaluation.

The network was trained for 200 epochs. To comprehensively evaluate the quality of interpolated frames, Peak Signal-to-Noise Ratio (PSNR) [23] and Structural Similarity Index (SSIM) [24] were adopted as the primary quantitative metrics. In addition, the per-frame inference time (Runtime) was reported to measure computational efficiency. All quantitative results are averaged over three independent runs to ensure reliability.

B. Ablation Study

To verify the effectiveness of the proposed GDAM and its two components, GCA and DSA, we conducted systematic ablation experiments. The quantitative results are summarized in Table I.

TABLE I
Ablation Study Results on Vimeo90K

Model	PSNR (dB)	SSIM	Runtime (ms)
RIFE	35.43	0.977	16
RIFE + GCA	35.83	0.978	19
RIFE + DSA	35.61	0.980	20
RIFE + GDAM (RIFE-AOV)	**35.88**	**0.980**	22

The ablation results demonstrate that introducing the GDAM attention mechanism into the baseline RIFE model effectively improves interpolation quality in complex motion scenarios. Specifically, incorporating the GCA module enhances channel-wise feature selection and increases PSNR by 0.40 dB (from 35.43 to 35.83), mainly improving pixel-level reconstruction fidelity. In contrast, integrating the DSA module improves SSIM by 0.003 (from 0.977 to 0.980), indicating better structural consistency and sharper motion boundaries.

When GCA and DSA are combined to form the complete GDAM module, the resulting RIFE-AOV model achieves the best overall performance, reaching 35.88 dB in PSNR and 0.980 in SSIM. These results validate the complementary nature of channel attention and directional spatial attention. Although GDAM introduces an additional computational overhead of approximately 6 ms, the improved interpolation quality justifies this moderate increase in runtime.

C. Comparison with Other Methods

To further evaluate the overall effectiveness of the proposed approach, RIFE-AOV is compared with several representative real-time or near real-time video frame interpolation methods, including CAIN [9], Super SloMo [8], SepConv [10], AdaCoF [25], EDSC [26], and the original RIFE. All methods were evaluated under identical experimental settings. The quantitative comparison is reported in Table II.

TABLE II
Comparison with State-of-the-Art Methods on Vimeo90K

Model	PSNR (dB)	SSIM	Runtime (ms)
CAIN	34.65	0.973	38
Super SloMo	34.64	0.974	62
SepConv	33.79	0.970	51
AdaCoF	34.27	0.971	34
EDSC	34.84	0.975	46
RIFE	35.43	0.977	16
RIFE-AOV (Ours)	**35.88**	**0.980**	22

As shown in Table II, the proposed RIFE-AOV achieves the best overall performance among all compared methods. In terms of interpolation quality, RIFE-AOV significantly outperforms both traditional and learning-based baselines, including the original RIFE, on PSNR and SSIM, demonstrating superior reconstruction fidelity and structural preservation.

Regarding computational efficiency, RIFE-AOV requires 22 ms per frame, which is slightly slower than the original RIFE (16 ms) but remains substantially faster than all other compared methods, the fastest of which is CAIN at 38 ms per frame. Overall, RIFE-AOV achieves a more favorable trade-off between interpolation quality and inference speed, validating the effectiveness of the proposed GDAM-based enhancements.

D. Experimental Results Analysis

Overall, the experimental results consistently demonstrate the effectiveness of the proposed RIFE-AOV across both accuracy-oriented and efficiency-oriented evaluations. From the ablation study, it can be observed that the proposed GDAM module provides stable and complementary improvements, where channel-wise attention primarily enhances pixel-level reconstruction accuracy, while directional spatial attention contributes more significantly to structural integrity and motion boundary preservation. Their combination yields the best overall performance with only moderate computational overhead.

In comparison with existing state-of-the-art methods, RIFE-AOV achieves a favorable balance between interpolation quality and inference speed. Although some methods adopt heavier network architectures to improve reconstruction accuracy, they suffer from substantially higher runtime, making them less suitable for real-time deployment. In contrast, RIFE-AOV maintains real-time performance while delivering superior PSNR and SSIM, highlighting its practical advantages in resource-constrained and latency-sensitive AOV scenarios.

These results indicate that explicitly enhancing optical flow estimation with lightweight attention mechanisms is an effective strategy for addressing the frame rate discontinuity problem caused by low-frame-rate triggering, without sacrificing computational efficiency.

V. Conclusions

This paper investigates the frame rate transition problem in Always-on Video systems, where switching from low-frame-rate standby mode to high-frame-rate recording introduces severe temporal discontinuity. Such frame rate transitions often lead to noticeable visual stuttering and motion detail loss at the early stage of event-triggered recording. To address this issue, we propose RIFE-AOV, an efficient real-time video frame interpolation method specifically designed for AOV frame rate switching scenarios.

Built upon the RIFE framework, RIFE-AOV enhances the optical flow estimation network by introducing a lightweight Global Directional Attention Mechanism into the core IFBlock. By jointly modeling global channel-wise feature importance and directional spatial structures, the proposed attention mechanism improves motion representation continuity and boundary modeling during frame rate transitions, while maintaining low computational complexity suitable for real-time deployment.

Extensive experimental results demonstrate that the proposed method achieves a favorable balance between interpolation quality and inference efficiency under frame rate transition conditions, making it well suited for practical AOV applications. The results further indicate that targeted attention enhancement within optical flow estimation is an effective strategy for mitigating temporal artifacts caused by frame rate switching in event-triggered video systems.

References

[1] G. L. Moepi and T. E. Mathonsi, "Smart surveillance systems: Trends, challenges and future directions," *Indonesian Journal of Computer Science*, vol. 14, no. 2, 2025.

[2] Z. Huang, T. Zhang, W. Heng, S. Li, and D. Chen, "RIFE: Real-time intermediate flow estimation for video frame interpolation," *arXiv preprint arXiv:2011.06294*, 2020.

[3] D. Kye, C. Roh, S. Ko, C. Eom, and J. Oh, "AceVFI: A comprehensive survey of advances in video frame interpolation," arXiv:2506.01061, 2025.

[4] Z. Shi, X. Liu, K. Shi, L. Dai, and J. Chen, "Video frame interpolation via generalized deformable convolution," *IEEE Trans. Multimedia*, vol. 24, 2022.

[5] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artificial Intelligence (IJCAI)*, pp. 674–679, 1981.

[6] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

[7] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, Mar. 2011.

[8] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super SloMo: High quality estimation of multiple intermediate frames for video interpolation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 9000–9008, 2018.

[9] W. Bao, W. Lai, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-aware video frame interpolation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3703–3712, 2019.

[10] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive convolution," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 670–679, 2017.

[11] X. Liu, Z. Gao, L. Chen, C. Shen, and A. van den Hengel, "Learning to interpolate via compactly supported separable convolution," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pp. 364–372, 2017.

[12] S. Niklaus and F. Liu, "Context-aware synthesis for video frame interpolation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1701–1710, 2018.

[13] M. Choi, S. Kim, M. Kim, and J. Lee, "Channel attention is all you need for video frame interpolation," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 34, no. 7, pp. 10663–10671, 2020.

[14] J. Kong, C. Jiang, and J. Yang, "IFRNet: Intermediate feature refine network for efficient frame interpolation," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pp. 1–10, 2021.

[15] Z. Shi, X. Zhang, Z. Wang, and H. Li, 'VFIformer: Video frame interpolation transformer for video enhancement," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 17684–17693, 2022.

[16] H. Park, J. Lee, and S. Lee, 'AMT: All-pairs multi-field transforms for efficient frame interpolation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 9801–9810, 2023.

[17] S. Kong, C. Lin, and K. Chen, 'Large motion video frame interpolation with multi-frame modeling," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pp. 11321–11330, 2023.

[18] Z. Huang, Y. Zhou, and J. Yang, 'Fast video frame interpolation via efficient motion representation," in *Proc. European Conf. Computer Vision (ECCV)*, pp. 395–411, 2022.

[19] J. Xu, L. Chen, and S. Wang, 'LiteVFI: Lightweight video frame interpolation for real-time applications," *IEEE Trans. Image Process.*, vol. 32, pp. 4217–4229, 2023.

[20] Y. Li, Q. Zhao, and X. Guo, 'Temporal-consistent real-time video frame interpolation with efficient attention," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 38, no. 5, pp. 4123–4131, 2024.

[21] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," arXiv:2112.05561, 2021.

[22] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *Int. J. Comput. Vision*, vol. 127, no. 8, pp. 1106–1125, Aug. 2019.

[23] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electron. Lett.*, vol. 44, no. 13, pp. 800–801, Jun. 2008.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[25] H. Lee, T. Kim, T. Y. Chung, D. Pak, Y. Ban, and S. Lee, "AdaCoF: Adaptive collaboration of flows for video frame interpolation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 5316–5325, 2020.

[26] X. Cheng and Z. Chen, "Multiple video frame interpolation via enhanced deformable separable convolution," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 2021, doi: 10.1109/TPAMI.2021.3100714.