Video-based Visible-Event Cross-modal Person Re-identification for Edge AI Surveillance Systems

Xinyun Zhang*, Zixiao Wang*, Yurui Kuang, Bei Yu
The Chinese University of Hong Kong

***Abstract*—Video-based cross-modal person re-identification (ReID) is a critical task for video surveillance and security systems, particularly in resource-constrained edge AI environments. While existing cross-modal ReID methods primarily focus on thermal-visible matching, event cameras, with their low power consumption, high temporal resolution, and sparse data representation, offer significant advantages for edge-based surveillance systems by reducing data processing overhead and enabling robust performance under challenging lighting conditions. In this paper, we introduce a novel task: video-based visible-event person re-identification (VE ReID), which aims to match identities across RGB and event camera modalities. To the best of our knowledge, this is the first work to systematically define and investigate this cross-modal task in the context of event-driven edge AI. Specifically, we curate evaluation benchmarks from existing RGB-event datasets and synthesize a new RGB-event dataset, explicitly adapting them to the cross-modal ReID setting to enable a more comprehensive evaluation of VE ReID. Extensive experiments reveal that existing cross-modal state-of-the-art (SOTA) methods fail to effectively address the unique challenges posed by event data, highlighting the importance of tailored solutions for this task. To this end, we propose a novel method that constructs auxiliary modalities using frequency information from RGB and event tracklets, aligning them effectively through a fine-grained metric learning loss. Our approach not only achieves significant accuracy improvements over existing methods but also demonstrates the potential of event cameras for efficient and scalable edge AI surveillance applications. All source code and benchmarks are publicly available at https://github.com/yxgnahz/ASPDAC26-Event-RGB-ReID.**

I. Introduction

Multi-modal person Re-ID plays a crucial role in surveillance systems, particularly in edge AI settings, by leveraging the strengths of diverse sensors to enhance performance under real-world constraints. Previous research has largely focused on the visible-infrared ReID setting (VI ReID), utilizing the advantages of infrared cameras for low-light conditions, as shown in Fig. 1(a). However, infrared cameras, while effective, often come with high power consumption and limited dynamic range, making them less suitable for resource-constrained edge devices.

Recently, biologically inspired sensors, such as event cameras, have emerged as a promising alternative for edge AI applications. Unlike traditional cameras, event cameras asynchronously capture changes in brightness at the pixel level, producing sparse event streams with sub-millisecond latency, a dynamic range exceeding 120 dB, and exceptional resilience to motion blur. These characteristics make them highly efficient in terms of power consumption, data sparsity, and robustness [1]–[4], offering significant advantages in edge-based surveillance scenarios, particularly on power-constrained or mobile platforms such as drones. Their low power consumption, high temporal resolution, and motion robustness are critical for real-time identity matching in tasks like drone-based aerial surveillance [5], where they can efficiently track fast-moving individuals while consuming minimal resources. These benefits position event cameras as a key sensor modality for edge AI surveillance systems.

* denotes equal contribution

The project is supported in part by Research Grants Council of Hong Kong SAR (No. CUHK14210723).

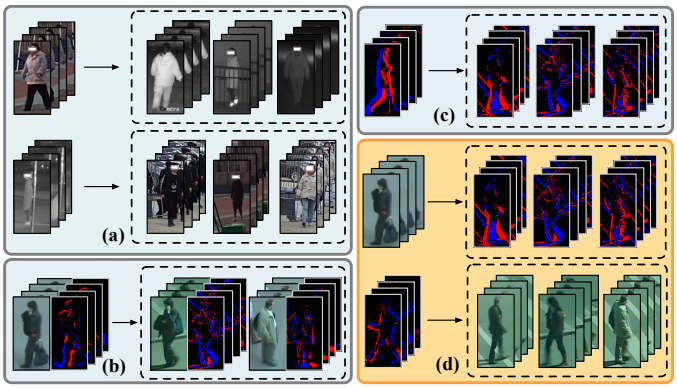

Fig. 1 An illustration of different ReID settings. (a) VI ReID [6], [7]; (b) Event data as a complement [8]; (c) Event-only ReID [9]; (d) VE ReID (Ours). Our work is the first to investigate the **cross-modal** visible-event setting for edge AI surveillance systems with event cameras.

Recognizing this potential, prior research [8], [9] has explored event cameras for ReID, focusing on two settings: a multi-modal setting, where event data complements RGB information in both the query and gallery sets [8] (see Fig. 1(b)), and a single-modality setting, which uses only event-based data [9] (see Fig. 1(c)). In contrast, this paper investigates a novel **cross-modal** setting for visible-event ReID (VE ReID), performing retrievals across the RGB and event modalities (see Fig. 1(d)). This task is highly relevant for edge AI surveillance systems where different cameras are deployed to fulfill complementary roles—for instance, RGB cameras at stationary locations and event cameras on power-constrained moving devices like drones or robots. This presents a more challenging domain due to the significant discrepancy between the visible and event data. The effectiveness of existing cross-modal Re-ID methods [6], [7], [10]–[16] in addressing this challenge remains an open research question.

Meanwhile, the scarcity of high-quality benchmarks poses another challenge to VE ReID. To address this, we first curate two benchmarks by adapting existing RGB-event versions of two widely-used ReID datasets—PRID-2011 [17] and MARS [18], originally provided by [8]—into a cross-modal setting. Furthermore, we also synthesize a new RGB-event dataset based on existing ReID dataset HITSZ-VCM [6], using the V2E simulator [19] following [8], to enable a more comprehensive evaluation.

We benchmark SOTA cross-modal VI Re-ID methods [6], [7], [10]–[16] on our VE ReID benchmarks. Experimental results show that none achieve satisfactory performance, underscoring the challenge of VE Re-ID. Analyzing the event data reveals two key challenges in aligning event and RGB video tracklets. First, event cameras capture sparse, binary intensity-change signals, primarily activating pixels at edges of moving identities without retaining color or dense intensity distributions, unlike RGB images (see Figure 2). These

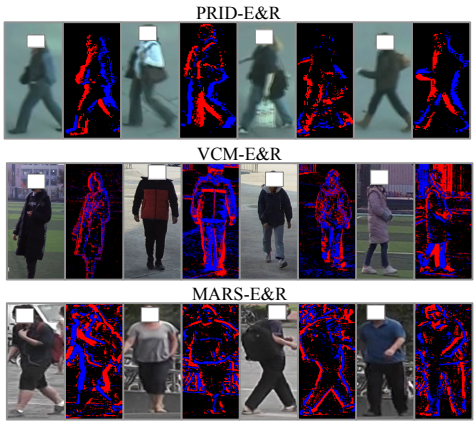

Fig. 2 Visualizations for the VE ReID benchmarks.

inherent differences cause significant discrepancies not only in spatial appearance but also in frequency domain characteristics. Specifically, RGB images predominantly encode rich low-frequency information due to smoothly varying intensity and color regions, whereas event data capture mainly high-frequency information corresponding to abrupt intensity transitions at edges and motion boundaries, as shown in Fig. 3. This feature creates significant discrepancies between RGB and event data. Second, unlike image-based Re-ID, each video tracklet comprises multiple frames, each potentially containing varied identity-related information. Simply pooling all frames may be suboptimal for the tracklet-level metric learning.

To address these challenges, we propose a novel learning framework that aligns RGB and event modalities using Fourier-based auxiliary modal learning and fine-grained frame-to-frame learning. Given the differing frequency domain distributions from the two sensors, we first construct auxiliary modalities by mixing the amplitude maps of the raw RGB and event images after Fourier transformation. We then align the auxiliary and raw modalities through a knowledge distillation-based consistency loss function. Additionally, we propose a new metric learning loss function that leverages the frame-to-frame distances to capture the fine-grained inter-tracklets distances.

We conduct extensive experiments on the three VE ReID benchmarks to validate our method's effectiveness. Results show that our approach significantly outperforms all previous SOTA cross-modal Re-ID methods, demonstrating its efficacy. Furthermore, testing on a traditional VI ReID benchmark confirms our method's superiority over SOTA methods, validating its generalization capability. Our contributions are as follows.

- To the best of our knowledge, we are the first to investigate the new cross-modal visible-event ReID task, specifically designed for edge AI surveillance systems with event cameras.
- We construct three VE ReID benchmarks through a combination of curating existing event-RGB datasets and synthesizing a new one, enabling comprehensive evaluations in cross-modal settings.
- We present a novel VE ReID framework, which incorporates Fourier-based auxiliary modal learning alongside a fine-grained, frame-level metric learning approach.
- Experimental results demonstrate the significant improvements of our proposed method over other SOTA cross-modal ReID methods in the VE ReID task.

II. Related Work

Cross-modal ReID. Most existing cross-modal ReID methods focus on visible-infrared settings. Image-based methods [10], [13]–[15], [20]–[37] can be mainly categorized into three types. Generative model-based methods [21]–[23] leverage GANs [38] to perform modality transformation. Meanwhile, part-based methods [10], [14], [24]–[26] mine the local region information for finer correspondences. Moreover, auxiliary modal-based methods [27]–[30] extract intermediate modalities from the RGB and infrared images for better alignment. XIV [30] first compresses RGB images into one single channel to create an X modal similar to thermal images. SMCL [27] uses heterogeneous image features to construct the auxiliary modality. MUN [28] develops a robust auxiliary modality containing rich semantic information. Our method also generates auxiliary modalities from the event and RGB input via spectrum amplitude mixup for enhanced cross-modal alignment. For video-based methods [6], [7], MITML [6] leverages a temporal memory refinement module to aggregate features from consecutive frames, while SAADG [7] proposes a style attack and defense mechanism to address the inter-modal discrepancy.

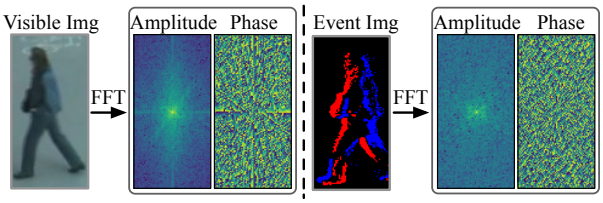

Fig. 3 An example demonstrating the spectral discrepancy between RGB and event images.

Event-based Vision. Existing works in the literature [2]–[4], [39]–[42] have investigated computer vision tasks for the bio-inspired event camera sensor. ESS [4] first reconstructs a grayscale image from event streams with off-the-shelf event-to-visible models [43] and aligns them with RGB images for semantic segmentation. RVT [2] proposes a recurrent vision transformer backbone for event-based object detection. Moreover, some works [8], [9] also investigate ReID with event cameras. SDCL [8] leverages event streams to enhance the RGB images in both query and gallery sets. [9] proposes a new benchmark and investigates event-to-event ReID. Different from them, our work focuses on a **cross-modal** setting, which has not been studied yet.

Fourier-based Methods. Previous studies [44]–[48] have incorporated Fourier transform into computer vision tasks. FDA [46] transforms the central region of the source image's amplitude spectrum to the target's for domain adaptation. Meanwhile, [44], [45] manipulate the amplitude map for tasks like image restoration and low-light enhancement. [48] leverages similar techniques for data augmentation in visible-infrared ReID. Inspired by these approaches, our work also utilizes amplitude spectrum maps; we generate auxiliary modalities from amplitude spectra and align them with the original modalities to effectively bridge the visible-event modality gap.

III. VE ReID Benchmarks

We propose three benchmarks for the cross-modal VE ReID task: PRID-E&R, MARS-E&R and VCM-E&R. The first two benchmarks are directly curated from the existing RGB-event variants of PRID-2011 [17] and MARS [18] provided by [8], adapting them specifically for the cross-modal ReID setting. To enable a more comprehensive evaluation, we synthesize event data streams from the RGB tracklets in HITSZ-VCM [6] and apply the same curation procedure, thereby obtaining our third evaluation benchmark, VCM-E&R. Following [8], [49], we use the V2E simulator [19]. Detailed configurations

Dataset	Type	Modal	#Identity		#Tracklet		#Frame	
			Train	Test	Train	Test	Train	Test
PRID-E&R	Real	Event	89	89	178	89	19077	9559
		RGB			178	89	19081	9826
VCM-E&R	Synthetic	Event	500	427	5300	2479	120981	56827
		RGB			5460	2569	126211	59460
MARS-E&R	Synthetic	Event	625	626	7588	3723	495269	258981
		RGB			7761	3664	504937	240050

TABLE I Detailed statistics of the VE ReID benchmarks. Type indicates whether the event streams are captured by real event cameras or synthetic.

are provided in the appendix. Note that the event streams of PRID-E&R are captured by real event cameras whereas those in MARS-E&R and VCM-E&R are synthesized using simulators. Therefore, our benchmarks include both real and synthetic data. Some visualizations and detailed statistics of the benchmarks are shown in Fig. 2 and TABLE I. We detail the data curation and evaluation metrics for VE ReID in the following sections.

A. VE ReID Data Curation

Cross-modal ReID Data Split. Existing RGB-event datasets, including our synthesized one, comprise a paired event stream for each RGB video tracklet. To adapt these datasets for the cross-modal ReID setting, we first perform a camera-level data split. Approximately half of the RGB cameras are designated as (pseudo) event cameras. For these (pseudo) event cameras, the event streams are retained while the RGB frames are discarded. Conversely, for the remaining RGB cameras, only the RGB frames are preserved, and the event streams are excluded. This camera-level split ensures the prevention of undesired information leakage between modalities. Additionally, we manually filter the data to ensure that each individual is captured by at least one RGB camera and one (pseudo) event camera. The train/test split is conducted at the individual level, following the original dataset's configuration. There are no overlapping identities in the train and test set.

Voxel-based Event Representation. Following [2], [41], we transform the event streams into voxel-based features to align with previous cross-modal ReID methods [6], [7]. An event stream comprises a set of event cases $\{\alpha_i\}_{i=1}^{N}$, where each event case $\alpha_i = [\omega_i, p_i, x_i, y_i]$ represents a brightness change p_i at location (x_i, y_i) at time ω_i. Our objective is to convert all the event cases into a dense 4-dimensional voxel tensor $\mathbf{E}$ with shape $(\Omega, 2, H, W)$. The transformation is formulated as follows:

$$\begin{aligned}&\mathbf{E}(\omega, p, x, y)\\ &= \max_{\alpha_i}\left\{\delta\left(\left\lfloor\frac{\omega-\omega_i}{\tau}\right\rceil\right)\delta(p-p_i)\delta(x-x_i, y-y_i)\right\},\end{aligned} \tag{1}$$

where $\delta(\cdot)$ denotes the Dirac function and τ signifies a time interval. Consequently, the event cases within each time interval are transformed into a dense frame $\mathbf{e} \in \mathbb{R}^{2\times H\times W}$. The two elements in the first dimension correspond to brightness increases and decreases, respectively. We set τ to be equivalent to the time interval between two frames in the RGB tracklet. Thus, through this event simulation and transformation, we obtain a tracklet of event frames that matches the length of the input RGB tracklet.

B. Evaluation Protocols

We follow the common evaluation protocols in cross-modal VI ReID [6], [7]. In the test set, the query and gallery sets consist of tracklets captured by different cameras. Additionally, we designate all tracklets in one modality as the query set and those in the other modality as the gallery set. To quantitatively assess the matching results, we employ the widely used Cumulative Matching Characteristic (CMC) curve and mean Average Precision (mAP). For a comprehensive evaluation, both retrieval modes—'visible-to-event (V2E)' and 'event-to-visible (E2V)'—are applied and reported.

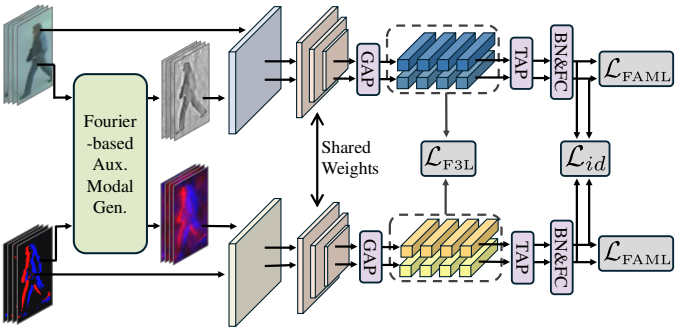

Fig. 4 The overview of the proposed framework. GAP and TAP denote global average pooling and temporal average pooling, respectively.

IV. Method

To mitigate the challenges for VE ReID, we propose a novel framework as depicted in Fig. 4. Motivated by the significant frequency domain discrepancy between RGB and event modalities, we introduce Fourier-based Auxiliary Modal Learning (**FAML**). Additionally, to better capture detailed identity cues across frames within tracklets, we further introduce Fine-grained Frame-to-frame Learning (**F3L**). These two components together effectively address the inherent modality gaps and enhance alignment performance.

A. Model Architecture

We utilize a ResNet architecture [50] as our backbone. In accordance with common practices for cross-modal ReID [6], [7], the two modalities employ separate initial convolutional blocks while sharing subsequent layers. We denote an RGB video tracklet as $\mathcal{T} = \{\mathbf{t}_i\}_{i=1}^{M}$, where $\mathbf{t} \in \mathbb{R}^{3\times H\times W}$ represents an individual RGB frame, and M is the length of the sampled sequence for each tracklet. Similarly, an event video tracklet is denoted as $\mathcal{E} = \{\mathbf{e}_i\}_{i=1}^{M}$, where $\mathbf{e} \in \mathbb{R}^{2\times H\times W}$ represents an event frame processed as described in Section III-A. The feature extraction for event and RGB frames can be denoted as:

$$\begin{aligned}f_{\text{event}}(\mathbf{e}) &= \text{Conv}_{2-5}(\text{Conv}_1^{\text{event}}(\mathbf{e})),\\ f_{\text{rgb}}(\mathbf{t}) &= \text{Conv}_{2-5}(\text{Conv}_1^{\text{rgb}}(\mathbf{t})).\end{aligned} \tag{2}$$

B. Fourier-based Auxiliary Modal Learning

To address the substantial frequency-domain gap between RGB and event modalities, FAML generates auxiliary modalities via amplitude mixup in the Fourier domain, which serve as bridges to effectively align the original modalities.

Fourier Transform. For any 2D image $\mathbf{x} \in \mathbb{R}^{H\times W}$ where H and W represent the image height and width, respectively, the Fourier transform is formulated as:

$$\mathcal{F}(\mathbf{x})(u, v) = \frac{1}{\sqrt{HW}}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\mathbf{x}(h, w)e^{-j2\pi\left(\frac{h}{H}u+\frac{w}{W}v\right)}. \tag{3}$$

Each element in $\mathcal{F}(\mathbf{x})$ is a complex number, and we can calculate its amplitude and phase components as follows:

$$\begin{aligned}\mathcal{A}(\mathbf{x})(u, v) &= \sqrt{\mathcal{R}^2(\mathbf{x})(u, v) + \mathcal{I}^2(\mathbf{x})(u, v)},\\ \mathcal{P}(\mathbf{x})(u, v) &= \arctan\left[\frac{\mathcal{I}(\mathbf{x})(u, v)}{\mathcal{R}(\mathbf{x})(u, v)}\right],\end{aligned} \tag{4}$$

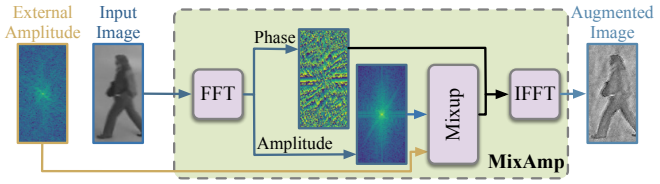

Fig. 5 An illustration of MixAmp(·).

where $\mathcal{R}(\mathbf{x})$ and $\mathcal{I}(\mathbf{x})$ denote the real and imaginary parts of $\mathcal{F}(\mathbf{x})$, respectively. We also denote the inverse Fourier transform as $\mathcal{F}^{-1}(\mathbf{x})$. Note that both the Fourier transform and its inverse can be efficiently implemented using the Fast Fourier Transform (FFT) algorithm [51].

Fourier-based Auxiliary Modal Generation. Event images capture pixel-wise motion information between consecutive frames, emphasizing moving boundaries between foreground and background (see Fig. 2). They also encode brightness intensity changes, losing original RGB color information in the process. These properties create significant frequency-domain discrepancies between event and RGB modalities, despite both modalities containing identical identity information, as shown in Fig. 3. To address this discrepancy, we construct auxiliary modalities that are frequency-wise closer to the opposing modality, facilitating improved cross-modal alignment. Previous studies, such as [46], [47], have validated that amplitude information is more domain- and style-relevant, whereas phase information is more semantically relevant. Inspired by this, we propose mixing amplitude information between event and RGB modalities to construct auxiliary modalities.

To start with, we define a MixAmp(·) function that mixes an external amplitude $\mathcal{A}'$ with an input image $\mathbf{x}$ as follows:

$$\begin{aligned}&\text{MixAmp}(\mathbf{x}, \mathcal{A}')\\ =&\mathcal{F}^{-1}((\lambda\mathcal{A}(\mathbf{x}) + (1-\lambda)\mathcal{A}') \cdot e^{-j\cdot\mathcal{P}(\mathbf{x})}),\end{aligned} \tag{5}$$

where $\lambda \sim \mathcal{U}(0,1)$ is the mixing ratio for mixup [52]. A schematic illustration is presented in Fig. 5. Subsequently, we calculate the amplitude information for both modalities. Since event images lack color information, we first convert the RGB image $\mathbf{t}$ into a grayscale format. In the context of event images, each event frame consists of two channels, $\mathbf{e}^{in}$ and $\mathbf{e}^{de}$, which signify density increases and decreases, respectively. We independently calculate the phase and amplitude for these channels. The amplitude of the event modality is represented by the average of the amplitudes from both channels. Then, the RGB-related auxiliary modal image is synthesized by integrating the event amplitude with the original RGB input as follows:

$$\begin{aligned}\mathbf{t}' &= \text{rgb2gray}(\mathbf{t}),\\ \tilde{\mathbf{t}} &= \text{MixAmp}(\mathbf{t}', \frac{1}{2}(\mathcal{A}(\mathbf{e}^{in}) + \mathcal{A}(\mathbf{e}^{de}))).\end{aligned} \tag{6}$$

Similarly, we can generate the auxiliary modality images for the event input as:

$$\begin{aligned}\tilde{\mathbf{e}}^{in} &= \text{MixAmp}(\mathbf{e}^{in}, \mathcal{A}(\mathbf{t}')),\\ \tilde{\mathbf{e}}^{de} &= \text{MixAmp}(\mathbf{e}^{de}, \mathcal{A}(\mathbf{t}')).\end{aligned} \tag{7}$$

The final auxiliary modal image for the event frame is is obtained by concatenating them, denoted as $\tilde{\mathbf{e}} = [\tilde{\mathbf{e}}^{in}; \tilde{\mathbf{e}}^{de}]$. During training, we randomly pair one event and one RGB tracklet, and the auxiliary modalities are generated based on the paired frames.

Cross-modal Alignment. Our objective is to align the generated auxiliary modalities with the original event and RGB modalities to enable the model to learn identity-related information while ignoring modal-specific noise. To accomplish this, we propose a distillation-based consistency loss applied to the predicted identity logits. Specifically,

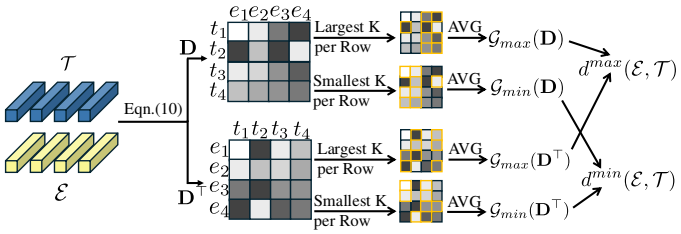

Fig. 6 A schematic illustration for computing the tracklet-to-tracklet distances.

the identity logits of an RGB tracklet are computed as follows:

$$\text{logits}(\mathcal{T}) = \text{Classifier}(\frac{1}{M}\sum_{i=1}^{M}(f_{\text{rgb}}(\mathbf{t}_i))), \tag{8}$$

. The logits for event tracklets $\text{logits}(\mathcal{E})$ can be obtained similarly. Then, the FAML consistency loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{FAML}} = &\text{JS}(\text{logits}(\tilde{\mathcal{T}})||\text{logits}(\mathcal{T}))\\ &+\text{JS}(\text{logits}(\tilde{\mathcal{E}})||\text{logits}(\mathcal{E})),\end{aligned} \tag{9}$$

where $\text{JS}(\cdot)$ represents Jenson-Shannon divergence.

C. Fine-grained Frame-to-frame Learning

Traditional video-based cross-modal Re-ID methods [6], [7] primarily utilize pooled tracklet features for metric learning. However, this approach is insufficient for capturing fine-grained tracklet-to-tracklet relations. To address this limitation, we propose a novel fine-grained triplet loss that leverages frame-to-frame interactions to model inter-tracklet distances.

Frame-to-frame distance. Given two tracklet $\mathcal{T}$ and $\mathcal{E}$, we first compute the frame-to-frame distance as follows:

$$\mathbf{D}_{ij}(\mathcal{T},\mathcal{E}) = \frac{f_{\text{rgb}}(\mathbf{t}_i)^\top f_{\text{event}}(\mathbf{e}_j)}{||f_{\text{rgb}}(\mathbf{t}_i)||_2||f_{\text{event}}(\mathbf{e}_j)||_2}. \tag{10}$$

Since we ensure the sequence length for each tracklet to be the same, $\mathbf{D}$ is an $M \times M$ square matrix. The next step involves aggregating the frame-to-frame distance matrix to calculate the tracklet-to-tracklet distance. To start with, we define two aggregation operations as follows:

$$\begin{aligned}\mathcal{G}_{\max}(\mathbf{D}) &= \frac{1}{M}\sum_{i=1}^{M}(\frac{1}{K}\sum_{j\in S_i}\mathbf{D}_{ij}),\\ \mathcal{G}_{\min}(\mathbf{D}) &= \frac{1}{M}\sum_{i=1}^{M}(\frac{1}{K}\sum_{j\in Q_i}\mathbf{D}_{ij}),\end{aligned} \tag{11}$$

where

$$\begin{aligned}S_i &= \text{largest}_K\left\{\mathbf{D}_{ij}|j=1\cdots M\right\},\\ Q_i &= \text{smallest}_K\left\{\mathbf{D}_{ij}|j=1\cdots M\right\}.\end{aligned} \tag{12}$$

In this context, let K denote a hyper-parameter that governs the aggregation range. We define two tracklet-to-tracklet distance metrics based on the two aggregation operations as:

$$d^{\max}(\mathcal{E},\mathcal{T}) = \frac{1}{2}\mathcal{G}_{\max}(\mathbf{D}(\mathcal{E},\mathcal{T})) + \frac{1}{2}\mathcal{G}_{\max}(\mathbf{D}^\top(\mathcal{E},\mathcal{T})), \tag{13}$$

and

$$d^{\min}(\mathcal{E},\mathcal{T}) = \frac{1}{2}\mathcal{G}_{\min}(\mathbf{D}(\mathcal{E},\mathcal{T})) + \frac{1}{2}\mathcal{G}_{\min}(\mathbf{D}^\top(\mathcal{E},\mathcal{T})). \tag{14}$$

An schematic illustration for this process is presented in Fig. 6. Intuitively, for each frame, we identify the top K largest or smallest frames in the other tracklet and compute the average over all the frames to derive two distinct tracklet-to-tracklet distance metrics.

Fine-grained Triplet Loss. Given a batch of tracklets $\mathcal{B}$ =

Method	Type	PRID-E&R						VCM-E&R						MARS-E&R					
		Visible-to-event			Event-to-visible			Visible-to-event			Event-to-visible			Visible-to-event			Event-to-visible		
		r=1	r=10	mAP	r=1	r=10	mAP	r=1	r=10	mAP	r=1	r=10	mAP	r=1	r=10	mAP	r=1	r=10	mAP
DDAG [10]	Image	28.09	70.79	40.89	29.21	67.41	41.93	75.87	88.36	64.16	77.41	88.79	65.53	75.04	91.04	57.01	69.31	87.40	56.56
AGW [11]	Image	24.72	66.29	38.30	24.72	66.29	37.57	79.88	89.57	67.22	81.08	90.96	69.09	78.47	92.82	61.25	74.72	90.34	62.06
MMN [12]	Image	40.45	78.65	51.99	34.83	78.65	48.75	70.22	85.52	59.31	72.21	86.37	61.21	70.19	87.59	50.39	64.63	84.47	49.62
MPANet [13]	Image	44.94	79.78	55.33	41.57	78.65	54.90	75.09	88.91	64.60	76.60	89.47	66.46	76.99	91.92	58.94	72.22	89.72	60.43
SAAI [14]	Image	50.56	85.39	62.19	47.19	84.27	59.42	77.50	91.24	68.47	79.27	90.00	69.99	78.66	93.42	62.69	74.67	91.14	64.31
DEEN [15]	Image	38.20	76.40	50.35	37.08	78.65	50.32	84.27	92.02	74.63	84.55	92.17	76.49	85.45	95.23	71.75	82.88	93.78	72.65
IDKL [16]	Image	35.96	79.78	50.42	47.19	80.90	57.77	64.27	82.72	53.33	59.90	82.65	50.83	64.77	87.12	46.18	58.57	84.76	47.34
MITML [6]	Video	26.97	66.29	41.28	29.21	65.17	40.61	46.70	75.93	47.41	45.44	73.38	46.78	75.10	91.42	57.85	70.93	87.75	58.54
SAADG [7]	Video	44.94	83.15	58.28	44.94	84.27	58.76	85.36	92.14	75.00	86.08	92.54	76.92	82.58	94.90	66.78	79.27	92.27	68.37
Ours	Video	**57.30**	**86.52**	**66.84**	**58.43**	**87.64**	**67.84**	**86.06**	**93.19**	**77.86**	**86.73**	**93.42**	**79.80**	**86.63**	**96.27**	**73.35**	**83.18**	**94.51**	**74.37**

TABLE II Comparison of our method and previous SOTA cross-modal ReID methods on VE ReID benchmarks.

Base	FAML	F3L	PRID-E&R		VCM-E&R		MARS-E&R	
			r=1	mAP	r=1	mAP	r=1	mAP
✓			40.45	51.47	79.15	68.57	78.99	65.01
✓	✓		47.19	59.49	86.40	77.04	83.61	71.84
✓		✓	51.69	63.26	83.87	75.11	83.40	71.36
✓	✓	✓	**57.87**	**67.34**	**86.40**	**78.83**	**84.91**	**73.86**

TABLE III Ablation study on the effectiveness of different modules. We report the average of E2V and V2E settings. See appendix for full results.

Aux Modal	Type	PRID-E&R		VCM-E&R		MARS-E&R	
		r=1	mAP	r=1	mAP	r=1	mAP
Grayscale	Img	40.45	53.80	81.22	71.68	82.04	69.85
Mixup	Img	37.64	50.48	83.73	74.68	82.08	70.13
PatchMix [53]	Img	39.33	52.63	83.40	73.53	80.84	68.54
MMN [12]	Feat	29.78	43.16	71.05	60.27	70.11	56.50
SMCL [27]	Feat	42.14	54.37	81.75	70.87	78.08	65.03
Fourier (Ours)	Img	**47.19**	**59.49**	**86.40**	**77.04**	**83.61**	**71.84**

TABLE IV Comparison with other auxiliary modals. We report the average of E2V and V2E settings. See appendix for full results.

$\{\mathcal{T}_1, \cdots, \mathcal{T}_L, \mathcal{E}_1, \cdots, \mathcal{E}_L\}$, our proposed fine-grained triplet loss can be defined as:

$$\mathcal{L}_{\text{F3L}} = \frac{1}{2L} \sum_{v_a \in \mathcal{B}} \max(0, \gamma + \max_{v_p \in \mathcal{Z}_p} d^{\max}(v_a, v_p) - \min_{v_n \in \mathcal{Z}_n} d^{\min}(v_a, v_n)), \tag{15}$$

where v_a represents any anchor tracklet (event or RGB) in the batch, γ is a hyper-parameter for the margin, and $\mathcal{Z}_p$ and $\mathcal{Z}_n$ represent the positive set and negative set, respectively. To explicitly mine hard samples, we employ different tracklet-to-tracklet distance metrics for positive and negative samples.

D. Loss Function

The total loss function consists of three parts: identity loss, metric learning loss, and consistency loss. The identity loss is a cross-entropy loss for ID classification, which can be formulated as:

$$\mathcal{L}_{Id}(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{v_i \in \mathcal{B}} \text{CE}(\text{logits}(v_i), y_i), \tag{16}$$

where y_i represents the id label for the i-th tracklet. Then, the total loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_{id}(\mathcal{B}) + \mathcal{L}_{id}(\tilde{\mathcal{B}}) + \mathcal{L}_{\text{F3L}}(\mathcal{B}) + \mathcal{L}_{\text{F3L}}(\tilde{\mathcal{B}}) + \mathcal{L}_{\text{FAML}}(\mathcal{B}, \tilde{\mathcal{B}}). \tag{17}$$

V. Experiments

A. Experimental Settings

Implementation Details. We utilize the ResNet50 model [50] as our backbone architecture, consistent with previous state-of-the-art studies [6], [7]. To ensure a fair comparison with existing video- and image-based approaches, we randomly select four identities per mini-batch, sampling two video tracklets for each identity captured by RGB and event cameras. For each tracklet, six frames are extracted by dividing the tracklet into six segments and selecting one frame per segment, following the methodology outlined in [6]. In alignment with [7], we conduct experiments over 200 epochs across all benchmarks. The initial learning rate is set to 0.1, with decays occurring at the 60th and 120th epochs at a decay ratio of 0.05. Our algorithm is implemented using PyTorch [54], and all experiments are executed on a single Nvidia 3090 GPU.

Benchmarks and Evaluation Metrics. For VE ReID, we conduct experiments on our proposed benchmarks: PRID-E&R, VCM-E&R, and MARS-E&R. The statistics for these benchmarks are detailed in TABLE I. We evaluate both event-to-visible (E2V) and visible-to-event (V2E) scenarios. Besides, we also assess the effectiveness of our method on the video-based infrared-RGB benchmark, HITSZ-VCM [6], following [6], [7]. HITSZ-VCM [6] includes 927 identities, with 500 used for training and the remaining 427 for testing. The training set comprises 6142 RGB and 4919 infrared tracklets, while the test set contains 5643 RGB and 5159 infrared tracklets.

B. Main Results

First, we benchmark previous SOTA cross-modal (IR-RGB) ReID methods [6], [7], [10]–[16] on our proposed event-RGB benchmarks. The results are shown in TABLE II. For image-based methods [10]–[16], we first extract the frame features and apply a temporal average pooling layer to obtain the tracklet feature, following [7]. Generally, all methods can not achieve satisfying performance compared with single-modality ReID (the best mAP is less than 80 on all these three benchmarks), demonstrating the challenges in VE ReID. The benchmark results show varied performance across methods and settings. SAADG [7] achieves the best performance on the VCM-E&R dataset. DEEN [15] shows superior results on the MARS-E&R dataset while remaining competitive on VCM-E&R. SAAI [14] outperforms other methods on the PRID-E&R benchmark. These findings highlight the specialized strengths of each method in enhancing cross-modal retrieval accuracy across varied scenarios.

Our method demonstrates superior performance across the three event-RGB benchmarks. On PRID-E&R, it exceeds the best baseline, SAAI [14], with margins of approximately 6% in mAP and 4% in rank-1 scores on average. Similarly, on VCM-E&R, our approach surpasses the leading video-based baseline, SAADG [7], by about 3% in mAP and 0.8% in rank-1 scores on average. On the largest benchmark, MARS-E&R, our method achieves improvements of 1.7% in mAP and 0.8% in rank-1 scores on average compared to the top baseline, DEEN [15].

Mixup	PRID-E&R		VCM-E&R		MARS-E&R	
	r=1	mAP	r=1	mAP	r=1	mAP
Fix-Full	46.07	57.97	84.84	74.95	80.04	67.32
Uni-Partial	**47.75**	58.59	85.67	76.82	83.20	71.37
Uni-Full (Ours)	47.19	**59.49**	**86.40**	**77.04**	**83.61**	**71.84**

TABLE V Ablation study on amplitude mixup methods. We report the average of E2V and V2E settings. See appendix for full results.

C. Ablation Study

Effectiveness of each module. We first conduct an ablation study to verify the effectiveness of the two proposed different modules. The results are shown in TABLE III. We start with a baseline model and progressively add FAML and F3L modules. The baseline model achieves rank-1/mAP scores of 40.45%/51.47%, 79.15%/68.57%, and 78.99%/65.01% on the three benchmarks, respectively. Adding the FAML module brings significant improvements across all datasets, with notable gains of +6.74%/+8.02%, +7.25%/+8.47%, and +4.62%/+6.83% in rank-1/mAP metrics. Similarly, incorporating the F3L module alone improves the baseline performance by +11.24%/+11.79%, +4.72%/+6.54%, and +4.41%/+6.35% in rank-1/mAP scores. When combining both modules, our full model achieves the best performance, demonstrating that FAML and F3L are complementary to each other.

Auxiliary Modal Comparisons. We compare our proposed Fourier-based auxiliary modal with other auxiliary modal approaches in cross-modal ReID [12], [27], [53]. Existing auxiliary modal generation methods can be categorized into two primary classes: input image-based [53] and feature-based [12], [27]. The former synthesizes auxiliary modalities by directly manipulating input images, such as implementing patch mixup [55], while the latter extracts auxiliary modals from latent representations. The results are presented in TABLE IV. Generally, leveraging simple image-based auxiliary modals, including grayscale images, standard mixup, and patch mixup [53] of RGB and event images, yields notable performance improvements across three benchmark datasets. In contrast, feature-based auxiliary modal methods [12], [27] demonstrate comparatively inferior performance. We hypothesize that this performance disparity stems from the substantial domain discrepancy between event and RGB data. Our proposed method outperforms all auxiliary modal baselines, substantiating the efficacy of leveraging spectral information from event and RGB modalities.

Effectiveness of mixup methods. We conduct an ablation study exploring different amplitude information mixing strategies. We compare our approach with two alternative methods: (1) direct swapping of complete amplitude spectra (Fix-Full), and (2) partial mixing confined to the central region (50% of height and width) with ratio sampling from a uniform distribution (Uni-Partial) [47]. In contrast, our proposed method samples a mixup ratio from a uniform distribution and subsequently merges the complete amplitude spectrum. The experimental results, presented in TABLE V, demonstrate that our method consistently outperforms both baseline approaches. This performance superiority suggests that our dynamic mixup technique, which operates on the full amplitude spectrum, can generate more diverse auxiliary modalities. Consequently, this approach enhances the robustness of the cross-modal representation learning framework.

Effectiveness of different choices of K. The hyperparameter K is critical in our proposed F3L module, governing the number of frames utilized during inter-tracklet distance computation. We conducted a comprehensive ablation study to systematically investigate the impact of varying K values, with results presented in TABLE VI. Empirical analysis reveals that $K = 2$ consistently demonstrates superior performance on most benchmarks. Consequently, we adopt $K = 2$ as the optimal hyperparameter for our proposed methodology.

K	PRID-E&R		VCM-E&R		MARS-E&R	
	r=1	mAP	r=1	mAP	r=1	mAP
1	50.56	62.60	81.95	73.40	83.29	71.25
2	**51.69**	63.26	**83.87**	**75.11**	**83.40**	**71.36**
3	50.00	62.25	83.51	74.76	82.46	70.07
4	51.13	**63.87**	82.90	74.22	83.12	71.01
5	50.56	62.95	83.19	74.56	83.13	71.07
6	47.19	59.61	82.04	73.37	82.93	70.64

TABLE VI Ablation study on K. We report the average of E2V and V2E settings. See appendix for full results.

Method	Visible-to-infrared			Infrared-to-visible		
	r=1	r=10	mAP	r=1	r=10	mAP
SAADG [7]	71.21	84.82	53.53	66.69	82.61	50.79
Ours	**71.29**	**86.04**	**55.39**	**67.98**	**84.21**	**53.59**

TABLE VII Results on infrared-RGB benchmarks.

D. Video-based Infrared-RGB ReID

To validate our method's transferability to other cross-modal settings, we evaluate it on the traditional video-based infrared-RGB ReID benchmark, HITSZ-VCM [6]. We compare our approach with the state-of-the-art method, SAADG [7], on this benchmark. For a fair comparison, we employ the same batch size (sampling four identities per batch), data augmentation techniques, and backbone architecture (ResNet50 with a modality-specific initial convolution block). The results, presented in TABLE VII, indicate that our method significantly outperforms its counterpart, achieving an average improvement of +2.33% in mAP. This demonstrates that our approach generalizes effectively to other cross-modal ReID settings beyond VE ReID.

VI. Conclusion

This paper addresses the novel problem of video-based visible-event person ReID (VE ReID), motivated by its relevance to edge AI surveillance systems. To facilitate research in this area, we construct three VE ReID benchmarks by curating existing RGB-event datasets and synthesizing a new benchmark tailored for cross-modal evaluation. Benchmarking state-of-the-art cross-modal VI ReID methods reveals their limitations in addressing the significant modality gap between RGB and event data, emphasizing the need for more effective solutions. To overcome these challenges, we propose a robust baseline that introduces Fourier-based auxiliary modality generation, which transforms RGB and event data into a shared frequency domain, effectively bridging the modality gap by capturing complementary information from both inputs. This auxiliary modality serves as an intermediary representation, enhancing the model's ability to align cross-modal features. Furthermore, we design a fine-grained frame-to-frame triplet loss to optimize tracklet-level representation learning, ensuring more discriminative and robust feature embeddings by focusing on temporal consistency and intra-tracklet relations. Our approach leverages these innovations to achieve superior cross-modal alignment and representation learning. Experimental results demonstrate that our method significantly outperforms prior methods on VE ReID benchmarks, achieving state-of-the-art performance while generalizing effectively to traditional VI ReID tasks. These findings not only validate the efficacy of our method but also highlight its potential to advance cross-modal ReID research and enable efficient, low-power identity recognition in edge AI applications.

REFERENCES

[1] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-video: Bringing modern computer vision to event cameras," in *CVPR*, 2019.

[2] M. Gehrig and D. Scaramuzza, "Recurrent vision transformers for object detection with event cameras," in *CVPR*, 2023.

[3] T. Kim, Y. Chae, H.-K. Jang, and K.-J. Yoon, "Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields," in *CVPR*, 2023.

[4] Z. Sun, N. Messikommer, D. Gehrig, and D. Scaramuzza, "Ess: Learning event-based semantic segmentation from still images," in *ECCV*. Springer, 2022.

[5] D. Falanga, S. Kim, and D. Scaramuzza, "How fast is too fast? the role of perception latency in high-speed sense and avoid," in *IEEE Robotics and Automation Letters*, 2019.

[6] X. Lin, J. Li, Z. Ma, H. Li, S. Li, K. Xu, G. Lu, and D. Zhang, "Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification," in *CVPR*, 2022.

[7] C. Zhou, J. Li, H. Li, G. Lu, Y. Xu, and M. Zhang, "Video-based visible-infrared person re-identification via style disturbance defense and dual interaction," in *ACM MM*, 2023.

[8] C. Cao, X. Fu, H. Liu, Y. Huang, K. Wang, J. Luo, and Z.-J. Zha, "Event-guided person re-identification via sparse-dense complementary learning," in *CVPR*, 2023.

[9] S. Ahmad, P. Morerio, and A. Del Bue, "Person re-identification without identification via event anonymization," in *ICCV*, 2023.

[10] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," in *ECCV*. Springer, 2020.

[11] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[12] Y. Zhang, Y. Yan, Y. Lu, and H. Wang, "Towards a unified middle modality learning for visible-infrared person re-identification," in *ACM MM*, 2021.

[13] Q. Wu, P. Dai, J. Chen, C.-W. Lin, Y. Wu, F. Huang, B. Zhong, and R. Ji, "Discover cross-modality nuances for visible-infrared person re-identification," in *CVPR*, 2021.

[14] X. Fang, Y. Yang, and Y. Fu, "Visible-infrared person re-identification via semantic alignment and affinity inference," in *ICCV*, 2023.

[15] Y. Zhang and H. Wang, "Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification," in *CVPR*, 2023.

[16] K. Ren and L. Zhang, "Implicit discriminative knowledge learning for visible-infrared person re-identification," in *CVPR*, 2024.

[17] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis: 17th Scandinavian Conference*. Springer, 2011.

[18] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *ECCV*, 2016.

[19] Y. Hu, S.-C. Liu, and T. Delbruck, "v2e: From video frames to realistic dvs events," in *CVPR*, 2021.

[20] M. Ye, W. Ruan, B. Du, and M. Z. Shou, "Channel augmented joint learning for visible-infrared recognition," in *ICCV*, 2021.

[21] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *CVPR*, 2020.

[22] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *CVPR*, 2019.

[23] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," in *ICCV*, 2019.

[24] H. Park, S. Lee, J. Lee, and B. Ham, "Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences," in *ICCV*, 2021.

[25] H. Sun, J. Liu, Z. Zhang, C. Wang, Y. Qu, Y. Xie, and L. Ma, "Not all pixels are matched: Dense contrastive learning for cross-modality person re-identification," in *ACM MM*, 2022.

[26] J. Wu, H. Liu, Y. Su, W. Shi, and H. Tang, "Learning concordant attention via target-aware alignment for visible-infrared person re-identification," in *ICCV*, 2023.

[27] Z. Wei, X. Yang, N. Wang, and X. Gao, "Syncretic modality collaborative learning for visible infrared person re-identification," in *ICCV*, 2021.

[28] H. Yu, X. Cheng, W. Peng, W. Liu, and G. Zhao, "Modality unifying network for visible-infrared person re-identification," in *ICCV*, 2023.

[29] Y. Zhang, S. Zhao, Y. Kang, and J. Shen, "Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification," in *ECCV*. Springer, 2022.

[30] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an x modality," in *AAAI*, 2020.

[31] X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, and L. Ma, "Farewell to mutual information: Variational distillation for cross-modal person re-identification," in *CVPR*, 2021.

[32] J. Feng, A. Wu, and W.-S. Zheng, "Shape-erased feature learning for visible-infrared person re-identification," in *CVPR*, 2023.

[33] Q. Zhang, C. Lai, J. Liu, N. Huang, and J. Han, "Fmcnet: Feature-level modality compensation for visible-infrared person re-identification," in *CVPR*, 2022.

[34] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, and N. Yu, "Cross-modality person re-identification with shared-specific feature transfer," in *CVPR*, 2020.

[35] J. Liu, Y. Sun, F. Zhu, H. Pei, Y. Yang, and W. Li, "Learning memory-augmented unidirectional metrics for cross-modality person re-identification," in *CVPR*, 2022.

[36] B. Yang, J. Chen, and M. Ye, "Towards grand unified representation learning for unsupervised visible-infrared person re-identification," in *ICCV*, 2023.

[37] X. Hao, S. Zhao, M. Ye, and J. Shen, "Cross-modality person re-identification via modality confusion and center aggregation," in *ICCV*, 2021.

[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," 2014.

[39] S. Schaefer, D. Gehrig, and D. Scaramuzza, "Aegnn: Asynchronous event-based graph neural networks," in *CVPR*, 2022.

[40] N. Messikommer, C. Fang, M. Gehrig, and D. Scaramuzza, "Data-driven feature tracking for event cameras," in *CVPR*, 2023.

[41] Y. Lu, Z. Wang, M. Liu, H. Wang, and L. Wang, "Learning spatial-temporal implicit neural representations for event-guided video super-resolution," in *CVPR*, 2023.

[42] X. Huang, Y. Zhang, and Z. Xiong, "Progressive spatio-temporal alignment for efficient event-based motion estimation," in *CVPR*, 2023.

[43] F. Paredes-Vallés and G. C. De Croon, "Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy," in *CVPR*, 2021.

[44] M. Zhou, J. Huang, C.-L. Guo, and C. Li, "Fourmer: An efficient global modeling paradigm for image restoration," in *International Conference on Machine Learning (ICML)*, 2023.

[45] C. Li, C.-L. Guo, M. Zhou, Z. Liang, S. Zhou, R. Feng, and C. C. Loy, "Embedding fourier for ultra-high-definition low-light image enhancement," in *ICLR*, 2023.

[46] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *CVPR*, 2020.

[47] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A fourier-based framework for domain generalization," in *CVPR*, 2021.

[48] X. Tan, Y. Chai, F. Chen, and H. Liu, "A fourier-based semantic augmentation for visible-thermal person re-identification," *IEEE Signal Processing Letters*, 2022.

[49] P. Duan, Z. W. Wang, X. Zhou, Y. Ma, and B. Shi, "Eventzoom: Learning to denoise and super resolve neuromorphic events," in *CVPR*, 2021.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[51] H. J. Nussbaumer and H. J. Nussbaumer, *The fast Fourier transform*. Springer, 1982.

[52] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[53] Z. Qian, Y. Lin, and B. Du, "Visible–infrared person re-identification via patch-mixed cross-modality learning," *Pattern Recognition*, 2025.

[54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[55] H. Zhang, "mixup: Beyond empirical risk minimization," *arXiv*, 2017.