ASP-DAC 2026
31st Asia and South Pacific Design Automation Conference

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

Xinyun Zhang*, Zixiao Wang*, Yurui Kuang, Bei Yu

Department of Computer Science & Engineering
The Chinese University of Hong Kong

Jan. 21, 2026

Outline

Introduction

What is Person Re-Identification (ReID)?

Definition:

- Retrieving a person of interest across non-overlapping camera views.
- Given a *query* image, find matches in a *gallery*.

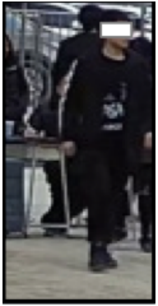
Standard ReID Setting

Core Challenges:

- Motion Blur
- Dependency on good lighting conditions.

The Evolution of ReID Settings

1. Visible (RGB)



Definition: Traditional ReID performed in daytime. It relies on single-modality visible light cameras, utilizing rich **color** and **texture** information.

2. Infrared (IR)

Definition: ReID performed in dark environments. It relies on thermal sensors, capturing heat signatures rather than color.

3. RGB-IR ReID

Definition: Cross-modality matching. The task is to match a query from one modality (RGB) to a gallery in another (IR), bridging the **modality gap**.

UAV Platform

Power

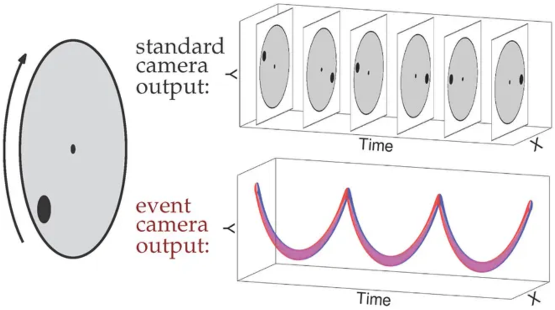
- Limited battery capacity.
- AI computation drains energy fast.

Motion Blur

- High speed causes image blur.
- Loss of visual details for ReID.

Goal: Efficient & Robust ReID on Edge Devices.

The Solution: Event Cameras



Asynchronous pixel response to motion.

Standard Cameras

- **High Data Rate:** Processes entire frame regardless of activity.
- **High Power:** Constant energy usage for redundant data.
- **Motion Blur:** Blurry images during fast motion.

Event Cameras

- **Ultra-Low Power:** Only active pixels consume energy (mW range).
- **High Speed:** Microsecond latency, no motion blur.
- **High Dynamic Range:** >120 dB.

Problem Definition: The VE-ReID Task

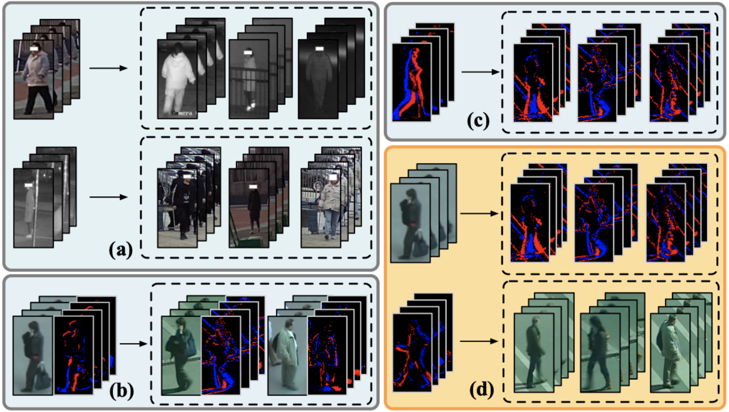


Fig. 1 An illustration of different ReID settings. (a) VI ReID [6], [7]; (b) Event data as a complement [8]; (c) Event-only ReID [9]; (d) VE ReID (Ours). Our work is the first to investigate the **cross-modal** visible-event setting for edge AI surveillance systems with event cameras.

Task Definition: Given an RGB/event query video tracklet, search the event/RGB tracklets belonging to the same person in the gallery set (cross-modal ReID).

Benchmarks and Challenges

Benchmark Construction: Event Data Representation

To align with video-based methods, raw event streams are transformed into **Voxel-based Features**.

- **Raw Input:** An event stream consists of event cases $\{\alpha_i\}_{i=1}^{N}$, where $\alpha_i = [\omega_i, p_i, x_i, y_i]$ (timestamp, polarity, coordinates).
- **Transformation:** Convert event cases into a dense 4D voxel tensor E with shape $(\omega, 2, H, W)$.

$$E(\omega, p, x, y) = \max_{\alpha_i} \left\{ \delta\left(\frac{\omega - \omega_i}{\tau}\right) \delta(p - p_i) \delta(x - x_i, y - y_i) \right\}$$

- τ: Time interval set to match the duration between two RGB frames.
- A dense frame $e \in \mathbb{R}^{2\times H\times W}$ (channels for brightness increase/decrease) for each time interval.
- This ensures the event tracklet length matches the RGB tracklet length.

Benchmark Construction: Data Generation Strategy

Due to the scarcity of large-scale real-world Event ReID datasets, we employ the **V2E**[1] **(Video to Events)** tool to synthesize high-quality event streams.

The V2E Tool

- **Function:** Converts standard RGB video frames into asynchronous event streams by simulating the internal mechanics of Dynamic Vision Sensors (DVS).
- **Process:**
  1. Takes high-frame-rate RGB videos as input.
  2. Calculates log-intensity changes between consecutive frames.
  3. Generates an event spike when the change exceeds a specific threshold C_{th}.

Application in this Work:

- We apply V2E to the **MARS** and **HITSZ-VCM** datasets.
- This allows us to leverage existing large-scale RGB annotations to create paired **Visible-Event** benchmarks (**MARS-E&R** and **VCM-E&R**).

[1] **Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck (2021). "v2e: From video frames to realistic DVS events".** In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1312–1321.

Benchmark Construction: Dataset Statistics

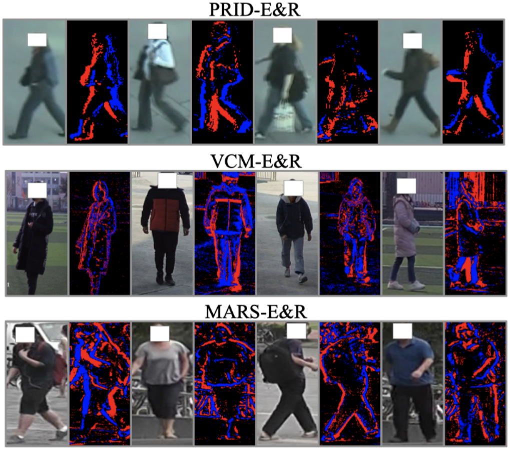
We curate three benchmarks: **PRID-E&R**, **MARS-E&R**, and **VCM-E&R**.

- **Data Sources:**
  - **PRID-E&R:** Real event cameras (curated from PRID-2011 variants).
  - **MARS-E&R:** Synthetic event data (sythesized from MARS using V2E simulator).
  - **VCM-E&R:** Synthetic event data (synthesized from HITSZ-VCM using V2E simulator).
- **Cross-modal Split:** Cameras are split into RGB-only and (Pseudo) Event-only to prevent information leakage.

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

The Challenges of VE ReID: Modality Discrepancy

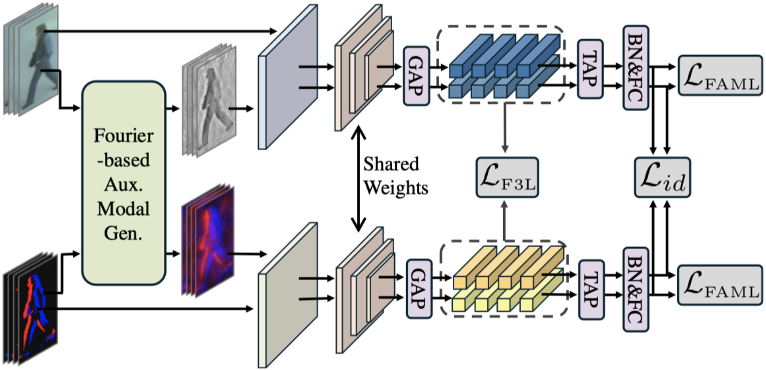
- **RGB:** Contains rich color and texture information; static intensity distribution.
- **Event:** Sparse, binary intensity changes; lacks color; primarily activates at edges of moving identities.



The Challenges of VE ReID: Video Tracklet Alignment

- Unlike image-based ReID, video tracklets contain multiple frames with varied identity information.
- Simple pooling (e.g., average pooling) is **suboptimal** for capturing fine-grained frame-to-frame correspondences.

Methodology



- **Backbone:** Two-stream ResNet50 (separate initial conv, shared subsequent layers).
- **Module 1: FAML** (Fourier-based Auxiliary Modal Learning) for bridging the event-RGB modality gap.
- **Module 2: F3L** (Fine-grained Frame-to-frame Learning) for capturing detailed temporal dynamics.

1. Motivation: The Frequency Gap

- **Observation:** As shown in **Fig. a**, RGB images are dominated by low-freq info (smooth intensity), while Event images capture high-freq info (edges/motion).
- **Challenge:** This spectral discrepancy makes direct alignment difficult.

2. Strategy: Auxiliary Modality Bridge

- **Why Fourier?**
  - *Amplitude* → Domain Style.
  - *Phase* → Semantic/ID Structure.
- **Solution:** We construct auxiliary modalities that sit *between* the domains.

3. Mechanism: MixAmp

- Swap **Amplitudes** to bridge the style gap while preserving **Phase** for ID consistency (**Fig. b**).

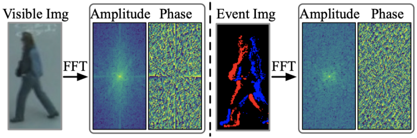


Fig. a: Spectral Discrepancy.

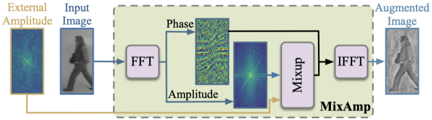


Fig. b: MixAmp Process.

Fine-grained Frame-to-frame Learning (F3L)

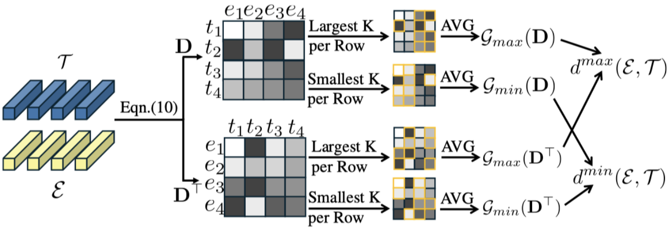
1. Limitation & Approach

- **Limitation:** Simple pooling (GAP/TAP) loses frame-level identity cues.
- **Solution:** Compute a dense **Frame-to-Frame Interaction Matrix (D_{ij})** between RGB (T) and Event (E) tracklets.

2. Aggregation Strategy

- Select Top-K *largest* ($\mathcal{G}_{max}$) and *smallest* ($\mathcal{G}_{min}$) values to mine hard samples.
- **Metrics:**

$$d_{max}(\mathcal{E},\mathcal{T}) = \frac{1}{2}\mathcal{G}_{max}(D) + \frac{1}{2}\mathcal{G}_{max}(D^\top)$$
$$d_{min}(\mathcal{E},\mathcal{T}) = \frac{1}{2}\mathcal{G}_{min}(D) + \frac{1}{2}\mathcal{G}_{min}(D^\top)$$



Optimization Objectives: Aligning Modalities & Tracks

1. FAML Consistency Loss ($\mathcal{L}_{FAML}$)

- **Goal:** Align the generated auxiliary modalities $(\tilde{T}, \tilde{E})$ with original inputs (T, E) to bridge the domain gap.
- **Formulation:** Uses Jensen-Shannon (JS) divergence on logits.

$$\mathcal{L}_{FAML} = JS(logits(\tilde{T})||logits(T)) + JS(logits(\tilde{E})||logits(E))$$

2. Fine-grained Triplet Loss ($\mathcal{L}_{F3L}$)

- **Goal:** Optimize tracklet representations via hard sample mining.
- **Formulation (Eq. 15):**

$$\mathcal{L}_{F3L} = \frac{1}{2|B|} \sum_{v_a \in B} \max\left(0, \gamma + \max_{v_p \in Z_p} d_{max}(v_a, v_p) - \min_{v_n \in Z_n} d_{min}(v_a, v_n)\right)$$

3. Total Loss

- Combines Identity Classification, Fine-grained Metric Learning, and Consistency:

$$\mathcal{L} = \mathcal{L}_{Id}(B, \tilde{B}) + \mathcal{L}_{F3L}(B, \tilde{B}) + \mathcal{L}_{FAML}(B, \tilde{B})$$

Experiments

Comparison With SOTA Methods

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**

Key Observations:

- **Superiority:** Outperforms SAAI (best image-based) by ∼**6% mAP** on PRID-E&R.
- **Video-based SOTA:** Surpasses SAADG by ∼**3% mAP** on VCM-E&R.
- **Consistency:** Achieves best performance across all three benchmarks in both V2E and E2V settings.

Does it work on traditional Visible-Infrared (VI) ReID?

To validate transferability, we evaluated on the **HITSZ-VCM** benchmark (Video-based IR-RGB).

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**

Conclusion:

- Our method achieves an average improvement of **+2.33% mAP** over the state-of-the-art SAADG.
- **Generalizability:** The proposed Fourier-based alignment and Fine-grained learning are effective not just for Event-RGB, but also for Thermal-RGB discrepancies.

Conclusion

Conclusion

Key Contributions

- **Pioneering Task Definition:**
  The **first work** to investigate Video-based Visible-Event (VE) ReID for Edge AI.
- **Benchmarks & Open Source:**
  Released three benchmarks (**PRID/MARS/VCM-E&R**) to facilitate research.
- **Methodology & Performance:**
  Proposed a Fourier-based framework (FAML & F3L) that achieves SOTA accuracy over existing methods.

Scan for Code & Datasets

THANK YOU!