

When Person Re-Identification Meets Event Camera: A Benchmark Dataset and An Attribute-guided Re-Identification Framework

Xiao Wang¹, Qian Zhu¹, Shujuan Wu¹, Bo Jiang¹, Shiliang Zhang^{2,3},
Yaowei Wang^{2,5}, Yonghong Tian^{2,3,4}, Bin Luo^{1*}

¹School of Computer Science and Technology, Anhui University, Hefei, China

²Peng Cheng Laboratory, Shenzhen, China

³School of Computer Science, Peking University, China

⁴School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, China

⁵Harbin Institute of Technology, Shenzhen, China

{zq542664, wushujuan0114}@163.com, {xiaowang, jiangbo, luobin}@ahu.edu.cn,
wangyw@pcl.ac.cn, {slzhang.jdl, yhtian}@pku.edu.cn

Abstract

github.com/Event-AHU/Neuromorphic_ReID

Recent researchers have proposed using event cameras for person re-identification (ReID) due to their promising performance and better balance in terms of privacy protection, event camera-based person ReID has attracted significant attention. Currently, mainstream event-based person ReID algorithms primarily focus on fusing visible light and event stream, as well as preserving privacy. Although significant progress has been made, these methods are typically trained and evaluated on small-scale or simulated event camera datasets, making it difficult to assess their real identification performance and generalization ability. To address the issue of data scarcity, this paper introduces a large-scale RGB-event based person ReID dataset, called EvReID. The dataset contains 118,988 image pairs and covers 1200 pedestrian identities, with data collected across multiple seasons, scenes, and lighting conditions. We also evaluate 15 state-of-the-art person ReID algorithms, laying a solid foundation for future research in terms of both data and benchmarking. Based on our newly constructed dataset, this paper further proposes a pedestrian attribute-guided contrastive learning framework to enhance feature learning for person re-identification, termed TriPro-ReID. This framework not only effectively explores the visual features from both RGB frames and event streams, but also fully utilizes pedestrian attributes as mid-level semantic features. Extensive experiments on the EvReID dataset and MARS datasets fully validated the effectiveness of our proposed RGB-Event person ReID framework. The benchmark dataset and source code will be released on https://github.com/Event-AHU/Neuromorphic_ReID.

1. Introduction

Person re-identification (ReID) is a critical research topic in the fields of computer vision and artificial intelligence. Its goal is to find pedestrians with the same ID as a given query sample from a set of candidate samples. It can be used in intelligent video surveillance, commercial and retail analysis, robotics and unmanned devices, etc. Most existing ReID algorithms are based on RGB cameras, making them highly susceptible to challenges such as illumination variations, motion blur, and privacy concerns.

To address the aforementioned issues, some researchers resort to the bio-inspired event cameras for person re-identification [2, 9, 24], due to their advantages in low energy consumption, high dynamic range, no motion blur, and spatial sparsity [12], as shown in Fig. 1. Specifically, Ahmad et al. [2] propose a person re-identification method through event anonymization, aiming to recognize individuals based on behavior and dynamic features without relying on identity information. Cao et al. [9] present an event-guided person re-identification method that leverages sparse-dense complementary learning, combining sparse event data and dense image features to enhance the robustness and accuracy of ReID across different scenarios. Li et al. [24] propose a video person re-identification method that enhances performance by leveraging cross-modality fusion of visual and event data, along with temporal collaboration to capture dynamic motion in challenging scenarios.

Despite these progress, current event-based person re-identification algorithms are still limited by the following

*Corresponding Author: Bin Luo

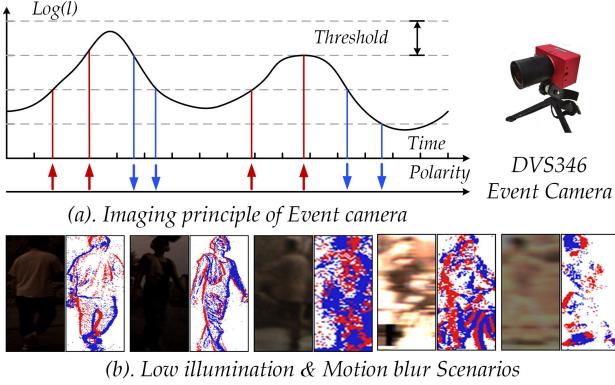


Figure 1. Imaging principle of event camera and comparison of RGB frames and event streams in challenging scenarios.

issues: 1). Existing person ReID algorithms are typically trained and evaluated on small-scale or simulated event camera datasets, making it difficult to assess their real identification performance and generalization ability. 2). Current ReID algorithms mainly focus on learning event stream features [2, 24] or fusing RGB and event features [9], but fail to consider the semantic information, such as pedestrian attributes *long hair, wearing glasses* [44], they can only achieve sub-optimal performance. Therefore, it is natural to raise the following question: “*How can we design a more effective and generalizable event-based person re-identification framework that leverages not only large-scale, real-world data but also incorporates rich semantic information such as pedestrian attributes?*”

Considering that the datasets for person re-identification based on event cameras are still scarce and limited in scale, for example, the Event ReID [2] dataset contains only 16,000 samples involving 33 pedestrians. In this paper, we first propose a new benchmark dataset for event stream-based person re-identification to bridge the data gap, termed EvReID. It contains 118,988 image pairs (7 times larger than the current real ReID dataset Event-ReID [2]) and covers 1200 pedestrian identities (36 times more than Event-ReID), with data collected across multiple seasons, scenes, and lighting conditions. We also evaluate 15 state-of-the-art person ReID algorithms, laying a solid foundation for future research in terms of both data and benchmarking. Some representative samples of our EvReID dataset are provided in Fig. 2.

Based on our newly proposed benchmark dataset, we further propose a new pedestrian attribute-guided contrastive learning framework to enhance feature learning for person re-identification, termed TriPro-ReID. This framework not only effectively explores the visual features from both RGB frames and event streams, but also fully utilizes pedestrian attributes as mid-level semantic features. Specifically, given

multi-modal input data, we first perform patching and projection to obtain RGB and event tokens, which are then fed into separate ViT backbone networks to learn spatio-temporal features. To facilitate feature interaction between the two modalities, we introduce a cross-modal prompt projector for cross-modal feature fusion. Meanwhile, we employ the VTFPAR++ [48] to predict pedestrian attributes from the input samples and use a text encoder to generate attribute semantic tokens. Subsequently, an attribute prompt injector is introduced to map these semantic tokens and fuse them with the visual features. We adopt widely-used metrics, including the triplet loss, ID loss, and the vision-attribute contrastive loss, to jointly guide the multi-modal feature learning for ReID. An overview of our proposed RGB-Event based person re-identification framework can be found in Fig. 3.

To sum up, the contributions of this paper can be summarized as the following three aspects:

- 1). We propose a large-scale benchmark dataset for RGB-Event based person re-identification, termed EvReID dataset, which contains 118,988 image pairs and covers 1200 pedestrian identities, with data collected across multiple seasons, scenes, and lighting conditions. 15 state-of-the-art (SOTA) person ReID algorithms are re-trained and evaluated on this dataset, which lays a solid foundation for future research in terms of both data and benchmarking.
- 2). We propose a pedestrian attribute guided contrastive learning framework for RGB-Event based person re-identification, termed TriPro-ReID. Both the semantic attribute and multi-modal visual features are exploited in our framework.
- 3). Extensive experiments conducted on two benchmark datasets (i.e., EvReID and MARS dataset) fully validated the effectiveness of our proposed RGB-Event person ReID framework.

The following of this paper is organized as follows: In Section 2, we review the related works on the event-based vision, person re-identification, and benchmark datasets. Then, we introduce our EvReID benchmark dataset in Section 3 and describe the key modules of our framework in Section 4. After that, we conduct the experiments and give both quality and quantity analysis in Section 5. Finally, we conclude this paper and propose possible research directions in Section 6.

2. Related Works

In this work, we propose to review the related works on Event-based Vision, Person Re-Identification, and Benchmark Datasets for Video-based Person Re-Identification. More works can be found in the following surveys [12].

2.1. Event-based Vision

Most computer vision systems rely on RGB modalities, which typically achieve superior performance. However, these systems tend to struggle in low-light environments. To address the shortcomings of RGB-based vision, numerous Event-based methods have been proposed, inspired by the unique pulse-like characteristics of events. In the re-identification field, Ahmad et al. [2] introduce the first Event-based person re-identification (ReID) dataset and develop an Event-anonymous network for privacy-preserving Event-based ReID. In the object detection field, RVT [13], GET [34], and SAST [35] achieve outstanding performance by leveraging a Transformer-based backbone, combined with an LSTM network and specialized attention mechanisms. MvHeat-DET [47] uses a novel heat conduction-based foundational model to fully harness the advantages of events, achieving an optimal balance between accuracy and speed. SpikeYOLO [32], SpikingYOLO [22], EMS-YOLO [39] employ an SNN network to model Event stream data and integrate it with YOLO for object detection. In tracking field, Wang et al. [46], Zhu et al. [60] utilize RGB and Event dual-modal data for tracking and introduce a new multimodal tracking dataset. Mamba-FETTrack [19] uses Mamba backbone networks to extract the features to realize more efficient tracking. However, even though Event-based methods address many of the limitations inherent in RGB cameras, RGB cameras still offer richer visual information, which Event cameras lack. This gap often results in Event-based methods achieving lower accuracy compared to their RGB counterparts. Therefore, we propose a dual-branch RGB-Event architecture to leverage the complementary strengths of both modalities, thereby improving performance.

2.2. Person Re-Identification

Person re-identification (ReID) [38] has received increasing attention, particularly in video-based scenarios where temporal coherence between frames is exploited for more robust matching. Early methods typically adopted CNNs as backbones, such as AP3D [14], which utilizes 3D convolutions to jointly model spatial-temporal cues, and TCLNet [16], which extracts complementary features through temporal modules. CTL [27], PSTA [49], and STRF [5] further enhance temporal modeling by integrating LSTM networks. Graph-based approaches have also been explored, with MGH [53] introducing hypergraph structures to capture high-order spatial-temporal relations, and Chen et al. [10] employing keypoint-based GCNs for refined motion representation. Meanwhile, Transformer-based architectures have shown strong potential for global context modeling. Works like DCCT [30], VDT [56], PFD [43], and SBM [7] adopt transformer backbones to improve representation quality. With the emergence of event cameras,

methods such as [1, 3, 4, 9] have explored event stream data to achieve high accuracy under challenging conditions while preserving privacy. Recently, the vision-language model CLIP [36] has demonstrated strong transferability, prompting its adoption in ReID. Li et al. [26] introduce a two-stage training strategy to align image features with textual supervision, while SVLL-ReID [40] generates diverse textual queries to enhance CLIP’s matching capacity. TF-CLIP [54] replaces static text with learned sequence representations, and CLIMB [55] combines CLIP with the Mamba architecture to balance accuracy and efficiency. Several works further incorporate external semantic knowledge. IDEA [51] generates structured captions for cross-modal consistency, and LATex [18] integrates attribute-based prompts to bridge the aerial-ground domain gap. Despite these advances, most existing methods either ignore attribute semantics or rely on heavy models for semantic enhancement. In contrast, we propose an attribute-guided dual-modality ReID framework that leverages lightweight attribute recognition to predict semantic descriptions, encodes them into text features via a text encoder, and fuses them with visual representations. A cross-modal prompting mechanism is further introduced to enable effective interaction and mutual enhancement between modalities, resulting in fine-grained, semantically enriched representations for robust cross-modal ReID.

2.3. Benchmark Datasets for Video-Based Person Re-Identification

The most widely used datasets for video-based person re-identification (ReID) include PRID-2011 [15], iLIDS-VID [42], MARS [57], DukeMTMC-VideoReID [52], and LS-VID [23]. PRID-2011 comprises 400 video sequences of 200 individuals captured by two static cameras, while iLIDS-VID contains 600 sequences of 300 pedestrians recorded under surveillance conditions from two disjoint camera views. As a large-scale benchmark, MARS is collected from six cameras, offering 17,503 tracklets for 1,261 identities along with 3,248 distractor sequences. DukeMTMC-VideoReID is a subset of the DukeMTMC [37] dataset, consisting of 702 identities for training, 702 for testing, and an additional 408 distractors. LS-VID further scales up video-based ReID with 3,772 annotated identities, offering more diverse visual scenarios. In addition to RGB-based datasets, the event modality has been introduced into the ReID domain. Ahmad et al. [2] proposed the first event-based ReID dataset, Event-ReID, which contains event sequences for 33 individuals. However, existing datasets are either limited to a single modality or constrained by scale, and to date, no real-world person ReID dataset simultaneously includes both RGB and event modalities under a unified setting. To address this gap, we introduce EvReID, a new benchmark designed for

dual-modality video-based ReID. It enables comprehensive evaluation of RGB, event, and RGB-Event fusion methods under consistent spatio-temporal conditions, and serves as a foundation for developing and benchmarking future multi-modal person ReID approaches.

3. EvReID Benchmark dataset

3.1. Protocols

To provide a good platform for the training and evaluation of RGB-Event person re-identification, we construct the EvReID benchmark dataset. When collecting data for the EvReID dataset, we obey the following protocols: 1). *Large Scale*: EvReID is the first video-based Re-ID dataset, which integrates both RGB and Event modalities. It includes 1,200 unique identities, captured over 118,988 frames, providing a comprehensive foundation for training and evaluation. 2). *Multiple Viewpoints*: We use a DVS346 Event camera to capture spatiotemporally aligned RGB-Event sample pairs. Specifically, by following the direction of the person’s movement and rotating the camera’s viewpoint, we obtain images of the same pedestrian from different viewpoints. 3). *Complex and Varied Scenes*: The dataset reflects real-world variability by incorporating pedestrian footage across different times of day, seasons, and weather conditions, ensuring it covers a wide range of scenarios that pedestrians encounter in practice. 4). *Bi-modal Complementarity*: We add 11 kinds of different noises on the RGB modal of the EvReID dataset, which include light change, motion blur, and adverse weather conditions, to validate complementary learning for enhanced bi-modality.

3.2. Data Collection and Statistical Analysis

As shown in Table 1, we compare several representative video-based person re-identification (ReID) datasets with our proposed EvReID dataset. Most existing benchmarks are based solely on the RGB modality, which performs well under well-lit conditions but struggles to capture rapid motion or subtle visual cues in low-light or nighttime environments. To address these limitations, recent efforts have attempted to simulate event-based reID by generating pseudo-event data from RGB frames. Notably, Ahmad et al. [2] proposed the first real-world Event-based ReID dataset, emphasizing privacy protection and robustness in complex scenes. However, their dataset contains only 33 identities, which significantly limits its scalability and suitability for large-scale ReID research. Moreover, many existing datasets are collected in overly constrained or homogeneous environments, limiting their generalization to real-world deployment scenarios.

In contrast, our proposed EvReID dataset is the first large-scale benchmark specifically designed for both Event-based and RGB-Event (RGE) person ReID tasks. As il-

lustrated in Fig. 2, compared with mainstream video-based person ReID datasets we show in the first row, EvReID provides diverse samples across multiple seasons, lighting conditions (including daytime and nighttime), and weather variations, making it significantly more representative of real-world applications. The dataset is collected using the DVS346 Event camera at a resolution of 346×260 and comprises 118,988 frame pairs from 1,200 distinct identities. Event streams are processed into Event frames and are spatiotemporally aligned with corresponding RGB frames, ensuring consistent temporal correspondence across modalities. To ensure fair evaluation, the dataset is split into non-overlapping training and test sets with a 70%/30% split. For evaluation, we adopt a single-shot setting, where the query and gallery sets are strictly separated.

3.3. Benchmark Baselines

To establish a comprehensive benchmark dataset for RGB-Event based person re-identification, we select 15 SOTA or representative methods for evaluation on our proposed dataset, including: 1). **CNN-based**: OSNet [58], TCLNet [16], AP3D [14], MGH [53], STMN [11], PSTA [49], GRL [29], BiCnet-TKS [17], SINet [6], and SDCL [9]. 2). **Transformer-based**: DCCT [30], CLIP-ReID [25], TF-CLIP [54], DeMo [50] and CLIMB-ReID [55].

To ensure a fair and unified evaluation across modalities, each baseline model is extended to support both RGB and Event stream inputs. This is achieved by duplicating the original single-stream architecture into two parallel branches, each responsible for processing one specific modality. Both branches share the same network structure and parameter configuration as the original design. After extracting features from the two modalities, the outputs are concatenated at the embedding level to generate a joint representation for person re-identification. This fusion strategy preserves the integrity of the original models and introduces no structural modifications, ensuring that performance comparisons remain focused on the influence of multimodal input rather than architectural differences.

4. Methodology

4.1. Preliminary: CLIP and CLIP-ReID

CLIP [36] consists of a visual encoder $\mathcal{V}(\cdot)$ and a text encoder $\mathcal{T}(\cdot)$, which are jointly trained using contrastive learning to project input images and texts into a shared representation space. Specifically, let $\{img_i, text_i\}_{i=1}^B$ denote a batch of B paired visual-language training samples, where $\{img_i\}$ is an image and $\{text_i\}$ is the corresponding textual description. CLIP employs the two encoders along with two linear projection layers to encode images and texts into corresponding feature embeddings. The similarity between

Table 1. Statistics of existing video-based person ReID datasets. **Seasons** indicates the diversity of data collection seasons, **Lighting** refers to the time periods of video capture, **Events-Reality** denotes whether the event data were acquired from real event sensors.

#Index	Dataset	Year	#IDs	#Images	Modality	Seasons	Lighting	Real Event
01	PRID-2011 [15]	2011	200	40,033	RGB	Single	Daytime	✗
02	SAIVT-Softbio [8]	2012	152	64,472	RGB	Single	Daytime	✗
03	iLIDS-VID [42]	2014	300	42,460	RGB	Single	Daytime	✗
04	HDA Person [33]	2014	53	2,976	RGB	Single	Daytime	✗
05	MARS [57]	2016	1,261	1,067,516	RGB	Single	Daytime	✗
06	DukeMTMC-VideoReID [52]	2018	1,404	815,420	RGB	Single	Daytime	✗
07	LS-VID [23]	2019	3,772	2,982,685	RGB	Single	Day, Night	✗
08	Event ReID [2]	2023	33	16,000	Event	Single	Daytime	✓
09	EvReID (Ours)	2025	1,200	118,988	RGB-Event	Multi	Day, Night	✓

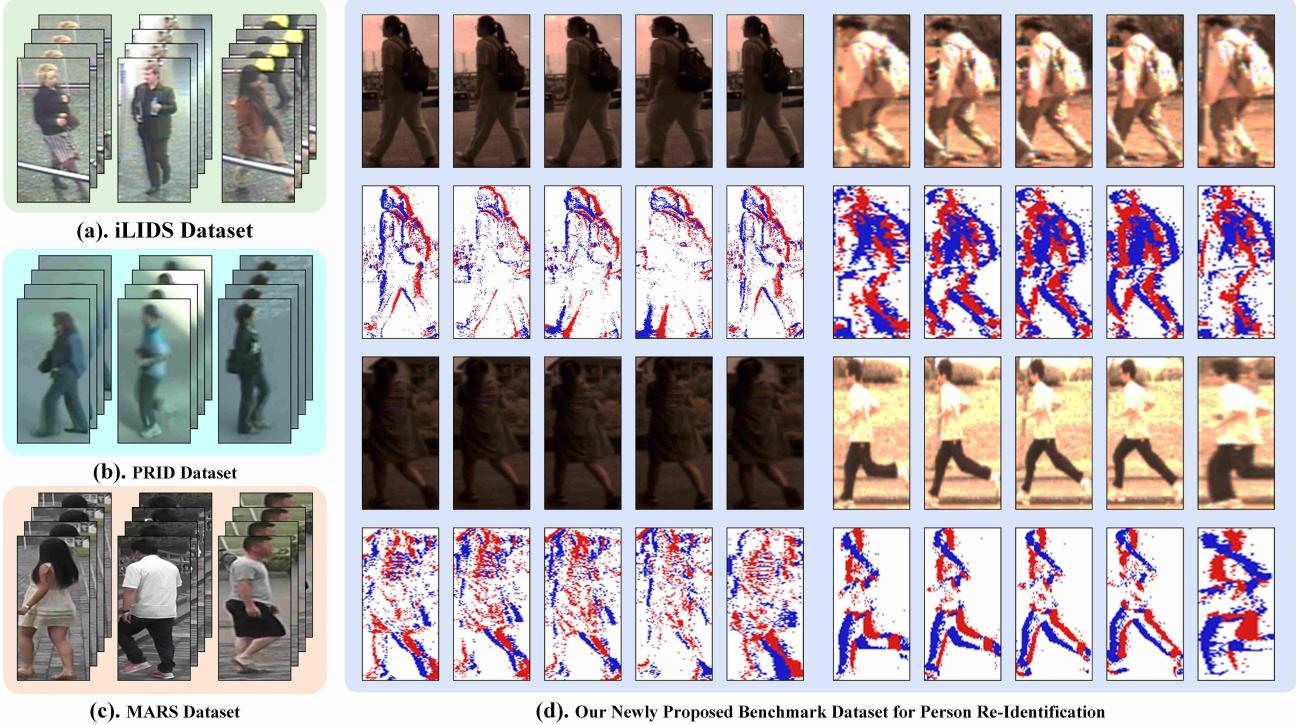


Figure 2. Comparison of existing video-based ReID datasets and our newly proposed EvReID dataset.

image and text features is computed as:

$$S(img_i, text_i) = \mathcal{J}(\mathcal{V}(img_i)) \cdot \mathcal{J}(\mathcal{T}(text_i)) \quad (1)$$

where \mathcal{J}_v and \mathcal{J}_t are linear layers that project features into a unified embedding space. Two contrastive losses, denoted as L_{v2t} and L_{t2v} , are employed to align visual and textual features during training. In downstream tasks, the textual input is typically formulated as prompts such as “A photo/video of a {class},” where “{}” is replaced with a specific class label (e.g., cat). However, in person ReID, identity annotations are merely represented as integer indices rather than meaningful text. To bridge this gap, CLIP-ReID [26] introduces learnable identity-specific tokens to

construct personalized textual prompts. These prompts take the form of “A photo of a $[X]_1, [X]_2, \dots, [X]_{Num}$ person”. During an additional training stage, the parameters of $\mathcal{V}(\cdot)$ and $\mathcal{T}(\cdot)$ are frozen, and the identity tokens are optimized using L_{v2t} and L_{t2v} to learn a discriminative but implicit textual representation for each identity. While CLIP-ReID achieves promising results on conventional RGB-based person ReID datasets, a simple dual-branch CLIP-ReID architecture proves insufficient when applied to RGB-Event dual-modality scenarios. It fails to fully leverage the complementary temporal and modality-specific cues provided by the Event stream. More critically, text prompts such as “A photo of a $[X]_1, [X]_2, \dots, [X]_{Num}$ person”, though

identity-aware, do not explicitly model or utilize the fine-grained pedestrian attribute information inherently present in video frames.

4.2. Overview

We propose a novel framework, **Triple-Prompting Re-identification** (TriPro-ReID), which consists of two core modules: Positive-Negative Attribute Prompting (PNAP) and Cross-Modal Prompting (CMP). The framework is optimized through a three-stage training strategy. In the first stage, we leverage the vanilla CLIP contrastive learning pipeline, initializing identity-aware context prompts for each person’s identity and associating them with corresponding IDs to obtain their textual representations. In the second stage, we fix the identity-aware context prompts and introduce visual prompts into the visual encoder, which helps align visual features with textual representations. Additionally, we introduce CMPs to bind the visual features from both RGB and Event modalities with the identity embeddings, and use the cross-modality prompt propagation to convey multimodal complementary information. This cross-modal alignment technique can effectively utilise complementary information from different modalities and extract generalized information from pre-trained CLIP models. In the third stage, we incorporate PNAPs, which are generated based on person attributes and encoded with the text encoder. These PNAPs are injected into the intermediate layers of the visual encoder to dynamically modulate the visual features with attribute-aware discrimination. The model is then supervised using ID and triplet losses to optimize it for person ReID tasks. The following sections will provide a detailed introduction to these modules.

4.3. Input Representation

Given the video-based person ReID training sets $\mathcal{D}_R = \{(x_i, y_i)\}_{i=1}^{N_s}$ and $\mathcal{D}_E = \{(x_i, y_i)\}_{i=1}^{N_s}$ in the RGB and Event modalities, respectively, where $y_i \in \{1, \dots, Y\}$ denotes the identity label and N_s is the total number of person tracklets in each modality, we aim to extract identity-specific sequence representations. Taking an identity y_i as an example, we first employ the pre-trained CLIP visual encoder to extract sequence-level features for all RGB and Event video sequences corresponding to y_i , denoted as \hat{F}_R and \hat{F}_E , respectively. Specifically, considering the RGB modality, we take an image sequence $\{I_{R_t}\}_{t=1}^T$ of T frames belonging to identity y_i as input. Each frame $I_{R_t} \in \mathbb{R}^{H \times W \times 3}$ is divided into N_p non-overlapping patches $I_{R_t,i} \in \mathbb{R}^{P^2 \times 3}$, where H and W denote the height and width of the image, and P is the patch size. The number of patches is given by $N_p = \frac{H}{P} \times \frac{W}{P}$. Each patch is projected into token embeddings via a linear projection layer $E_{\text{emb}} \in \mathbb{R}^{3P^2 \times D}$, where D is the token dimension. Additionally, a learnable class token [CLS] for each frame,

denoted as $CLS_{R_t} \in \mathbb{R}^D$, is prepended to the sequence of patch embeddings. The resulting input to the visual encoder at frame t is formulated as:

$$Z_{R_t}^{\text{Patch}} = [CLS_t, E_{\text{emb}}I_{t,1}, \dots, E_{\text{emb}}I_{t,N_p}] + e^{\text{sp}} \quad (2)$$

where e^{sp} denotes the spatial positional embedding. These embeddings are subsequently processed by the CLIP visual encoder. The resulting frame-level feature representation is:

$$\hat{F}_R = \mathcal{V}(Z_{R_t}^{\text{Patch}}) \quad (3)$$

where $\mathcal{V}(\cdot)$ denotes the CLIP visual encoder. To incorporate identity-level textual guidance, we construct an ID-specific prompt: “A photo of a $[X]_1, [X]_2, \dots, [X]_{\text{Num}}$ person”, where each $[X]_{y_i}$ ($y_i \in \{1, \dots, \text{Num}\}$) is a learnable token with the same dimension as the word embedding and is independently learned for each identity. The textual prompt is fed into the CLIP text encoder $\mathcal{T}(\cdot)$ to obtain the identity-specific text feature:

$$T_{y_i} = \mathcal{T}(\text{text}_{y_i}) \quad (4)$$

where text_{y_i} is the generated textual description for identity y_i .

4.4. Attribute Guided Prompt Learning

We design a dual-stream network architecture for RGB-Event person ReID, which leverages CLIP-ReID as our baseline and introduces two main modules to enhance feature fusion between RGB and Event modalities and explore attribute semantic information for robust person ReID.

- **Positive-Negative Attribute Prompts.** To further enhance the discriminative capacity of identity representations with fine-grained semantic cues, we propose a *Positive-Negative Attribute Prompting* (PNAP) mechanism. This design is inspired by the learnable context prompt strategy in [21, 59], and is further motivated by [51], which demonstrates the benefits of incorporating attribute-level semantic priors for multi-modal object re-identification. Specifically, we employ a pre-trained pedestrian attribute recognition model [48], denoted as $\mathcal{PAR}(\cdot)$, to predict attribute labels for a given input sequence. Based on the predicted results, we construct a *positive attribute prompt* A^{pos} , which contains all the predicted attributes (e.g., “Male, Jacket, Bald”), and a *negative attribute prompt* A^{neg} by negating the unpredicted attributes (e.g., “Not Female, Not Short Sleeves, Not Long Hair”). Both types of prompts are encoded into textual embeddings using the frozen CLIP text encoder $\mathcal{T}(\cdot)$. Following the prompt tuning paradigm of CoOp [59], we further introduce a set of learnable context vectors: P^{pos} for positive prompts and P^{neg} for negative ones. These are used to adapt the attribute prompts to the target person ReID task. The enriched prompts are then passed through a fully connected (FC) layer and concatenated with the visual features

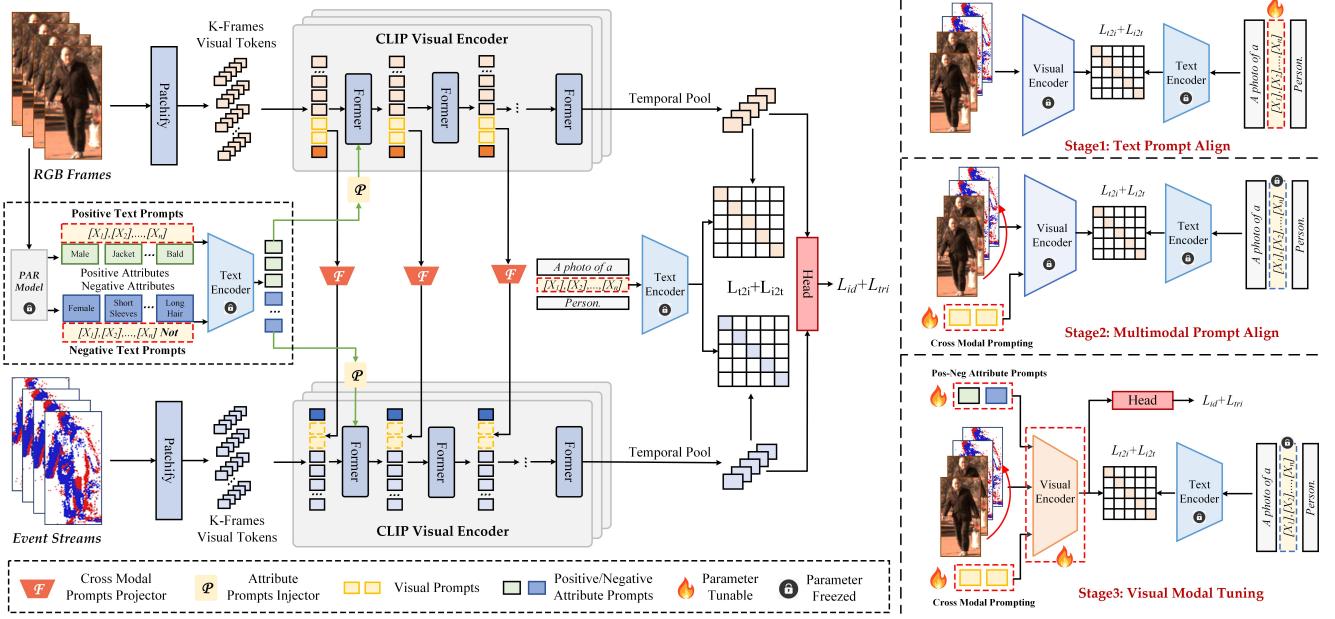


Figure 3. An illustration of our proposed TriProReId framework. The left part illustrates the overall architecture of the RGB-Event person re-identification framework, consisting of dual-stream backbones and multi-level fusion modules. The right part presents the three-stage training strategy, including Text Prompt Align, Multimodal Prompt Align, and Visual Modal Tuning.

\hat{F}_R and \hat{F}_E to yield prompt-enhanced identity representations F_R^0 and F_E^0 . The overall process can be summarized as:

$$\mathcal{A} = \text{PAR}(I_R) \quad (5)$$

$$PNAP = \mathcal{T}(\mathcal{A}^{\text{pos}}, P^{\text{pos}}) \cup \mathcal{T}(\mathcal{A}^{\text{neg}}, P^{\text{neg}}) \quad (6)$$

$$F_R^0 = \text{Con}(\hat{F}_R, \text{FC}(PNAP)) \quad (7)$$

$$F_E^0 = \text{Con}(\hat{F}_E, \text{FC}(PNAP)) \quad (8)$$

Here, $\text{PAR}(\cdot)$ is the attribute recognition model, $\mathcal{T}(\cdot)$ denotes the CLIP text encoder, and $\text{Con}(\cdot)$ represents the feature concatenation operation. F_R^0 and F_E^0 serve as the prompt-augmented input representations for the RGB and Event branches, respectively.

• **Cross Modal Prompts.** Inspired by the Visual Prompt Tuning paradigm [20], we design a Cross Modal Prompting (CMP) mechanism to facilitate bidirectional interaction between RGB and Event modalities at the feature level, which introduces a set of learnable prompts to enable early-stage alignment across modalities. Concretely, we initialize a set of modality-specific prompt tokens CMP_R^0 for the RGB branch and project them into the Event branch via an FC layer to obtain CMP_E^0 . These prompts are concatenated with the visual features F_R^0 and F_E^0 , resulting in the prompt-

jected features F_R^1 and F_E^1 , respectively:

$$F_R^1 = \text{Con}(F_R^0, CMP_R^0) \quad (9)$$

$$CMP_E^0 = \text{FC}(CMP_R^0) \quad (10)$$

$$F_E^1 = \text{Con}(F_E^0, CMP_E^0) \quad (11)$$

Here, F_R^0 and F_E^0 denote the visual features at the input (i.e., layer 0) of the transformer for the RGB and Event branches, respectively, while CMP_R^0 and CMP_E^0 represent the initial Cross Modal Prompt tokens for each modality. As the features propagate through the transformer layers, the RGB-side CMPs are iteratively updated to CMP_R^1, \dots, CMP_R^k , where the superscript k indicates the corresponding transformer layer. At each layer, the CMPs in the Event branch are synchronously replaced with the transformed RGB-side CMPs from the same layer, enabling continuous and guided feature fusion between the two modalities throughout the network.

4.5. Training Strategy

To effectively bridge the modality gap and enhance discriminative representation learning for RGB-Event person re-identification, we propose a three-stage training strategy, progressively aligning text, vision, and multimodal cues. The training stages are illustrated in Fig. 3 and described as follows.

• **Stage 1: Text Prompt Alignment.** In the first stage, we follow the paradigm of CLIP-ReID and introduce ID-specific learnable prompts to model ambiguous identity

related semantics. For each identity, we learn a set of n text tokens $[X]_1, [X]_2, \dots, [X]_n$, which are concatenated to form a textual description in the format of “A photo of a $[X]_1, [X]_2, \dots, [X]_n$ person”. These prompts are optimized while keeping both the visual encoder $\mathcal{V}(\cdot)$ and text encoder $\mathcal{T}(\cdot)$ frozen. The optimization is guided by two symmetric contrastive losses, vision-to-text (L_{v2t}) and text-to-vision (L_{t2v}), which encourage cross-modal alignment between video features and the learnable textual embeddings. Formally, given a mini-batch of size $|B|$, the overall objective is defined as:

$$L_{v2t} = -\frac{1}{|B|} \sum_{i=1}^{|B|} \log \frac{\exp(v_i^\top t_i / \tau)}{\sum_{j=1}^{|B|} \exp(v_i^\top t_j / \tau)} \quad (12)$$

$$L_{t2v} = -\frac{1}{|B|} \sum_{i=1}^{|B|} \log \frac{\exp(t_i^\top v_i / \tau)}{\sum_{j=1}^{|B|} \exp(t_i^\top v_j / \tau)} \quad (13)$$

$$L_{stage1} = L_{v2t} + L_{t2v} \quad (14)$$

where v_i and t_i represent the visual and text features of the i -th sample, τ is a temperature hyperparameter, L_{v2t} and L_{t2v} encourage the visual features to match their paired textual descriptions in the embedding space. This stage helps to warm up the model with fundamental visual-language alignment before introducing multimodal and fine-grained supervision.

• Stage 2: Multimodal Prompt Alignment. To integrate complementary information from both RGB and Event streams, we introduce learnable CMP in the second stage. These prompts serve as a bridge between modalities, enabling the model to capture and fuse heterogeneous visual patterns from both input types. Specifically, we concatenate the CMP tokens with the patch embeddings of RGB and Event frames, forming a unified multimodal token sequence. This sequence is then input to the frozen visual encoder $\mathcal{V}(\cdot)$, allowing the CMPs to guide the fusion process during token interactions. The text encoder $\mathcal{T}(\cdot)$ remains frozen, and we continue to apply the contrastive loss:

$$L_{stage2} = L_{v2t} + L_{t2v} \quad (15)$$

This stage focuses purely on visual modality fusion, enhancing the joint representation of RGB and Event data through CMP while retaining language guidance established in Stage 1.

• Stage 3: Visual-Modal Tuning with Attribute Prompts. In the final stage, we fine-tune the visual encoder by injecting PNAP to enhance attribute-level discrimination. These prompts are generated based on pre-estimated person attributes (e.g., “Male”, “Bald”, “Not Long Hair”) and encoded using the text encoder $\mathcal{T}(\cdot)$. The PNAP tokens are injected into intermediate layers of the visual transformer via a dedicated Attribute Prompt Injector, allowing the model to

Algorithm 1: Training Procedure of TriPro-ReID

Input: Train set \mathcal{D} , attributes for each identity, and CLIP encoders \mathcal{V} and \mathcal{T}

1 Stage 1: Text Prompt Alignment;
2 Freeze \mathcal{V} and \mathcal{T} ; Update ID-specific learnable text prompts;
3 for each batch $(I_{R,t}, I_{E,t}, y)$ in \mathcal{D} **do**
4 Concatenate the multimodal visual features in channel dimension, and apply temporal pooling to obtain final representations;
5 Compute contrastive loss $L_{v2t} + L_{t2v}$;

6 Stage 2: Multimodal Prompt Alignment;
7 Keep \mathcal{V} and \mathcal{T} frozen; initialize cross-modal prompts (CMPs);
8 for each batch $(I_{R,t}, I_{E,t}, y)$ in \mathcal{D} **do**
9 Inject CMP tokens into patch embeddings;
10 Concatenate and input into \mathcal{V} to obtain joint features;
11 Compute $L_{v2t} + L_{t2v}$;

12 Stage 3: Visual-Modal Fine-Tuning with Attribute Prompts;
13 Unfreeze \mathcal{V} ; Encode PNAPs via \mathcal{T} ;
14 for each batch $(I_{R,t}, I_{E,t}, y)$ in \mathcal{D} **do**
15 Inject PNAPs into \mathcal{V} via attribute prompter;
16 Concatenate CMP, PNAP, and patch embeddings;
17 Compute full loss: $L_{id} + L_{tri} + L_{v2t} + L_{t2v}$;

modulate visual features dynamically according to semantic cues. Additionally, both CMP and standard learnable visual prompts are preserved and fused at this stage, enabling simultaneous attribute awareness and modality integration. After temporal pooling over the fused visual tokens, the final feature is supervised by a classification head using ID and triplet losses:

$$L_{id} = \frac{1}{|B|} \sum_{i=1}^{|B|} \text{CrossEntropy}(v_i, y_i) \quad (16)$$

$$L_{tri} = \frac{1}{|B|} \sum_{i=1}^{|B|} \max(0, \|v_i - v_+\|_2^2 - \|v_i - v_-\|_2^2 + \alpha) \quad (17)$$

$$L_{stage3} = L_{id} + L_{tri} + L_{v2t} + L_{t2v} \quad (18)$$

where $|B|$ represents the batch size, v_i denotes the feature vector of the i -th sample, y_i is the true identity label, v_+ and v_- are the feature vectors of the positive and negative samples, respectively, $\|\cdot\|_2$ denotes the L2 norm (Euclidean distance), and α is the margin used in the triplet loss to en-

sure sufficient distance between positive and negative samples. The terms L_{id} , L_{tri} , L_{v2t} , and L_{t2v} together define the overall loss function, where L_{id} enforces identity classification, L_{tri} encourages correct relative distances between features, and L_{v2t} and L_{t2v} ensure cross-modal alignment between visual and textual features.

The whole training process of the proposed TriPro-ReID is summarized in Algorithm 1. We first leverage learnable contextual prompts to mine and store the hidden states of the pre-trained visual encoder and text encoder, allowing CLIP to retain its own advantages. During the second stage, the cross modal prompts are utilized to learn effective RGB-Event fusion. In the final stage, the positive-negative attribute prompts are joined to enhance attribute-aware discrimination.

5. Experiments

5.1. Datasets and Evaluation Metric

To evaluate the performance, we conduct a comprehensive benchmark of 10 pedestrian attribute re-identification methods, representing the most important models in the field of pedestrian attribute re-identification. Since there is no available RGB-Event person ReID dataset, Cao et al. [9] generate events from MARS [57]. We adopt this simulated dual-modal MARS RGB-Event dataset (MARS*) and our proposed EvReID dataset for evaluation. Following previous works, we adopt the mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank- K ($K = 1, 5, 10$) as our evaluation metrics.

5.2. Implementation Details

We use CLIP-ReID [26] as our baseline method, and the two modalities adopt independent backbone parameters. Our method consists of three training stages. In the first stage, we only optimize the ID-related context prompts. In the second stage, we introduce 20 randomly initialized cross-modal visual prompts into the RGB branch and generate Event-modality visual prompts via a projection layer. During this stage, only the cross-modal visual prompts and the projection layer are optimized. In the third stage, the text encoder and text context prompts remain frozen. We introduce learnable attribute context prompts and PNAP, and optimize them along with CMP and the visual backbone. SGD is used as the optimizer in the first two stages, while AdamW [31] is employed in the final stage. The learning rates are set to 3.5×10^{-3} , 3.5×10^{-3} , and $5e-6$, respectively. All training is conducted on an NVIDIA RTX 3090 GPU, with batch sizes of 64, 64, and 8, respectively. More details can be found in our source code.

5.3. Comparison with Other SOTA Algorithms

- **Result on EvReID Dataset.** As shown in Table 2, our model achieves impressive mAP, Rank-1, Rank-5, and Rank-10 scores of 69.3/88.6/94.3/95.4, respectively, significantly surpassing existing baseline methods. This substantial performance gain highlights the effectiveness of our dual-branch architecture, which fully exploits the complementary strengths of both RGB and Event modalities. Unlike prior ReID methods, including CNN-based, graph-based, and Transformer-based approaches, our model incorporates a hierarchical cross-modal prompting mechanism that seamlessly integrates multi-modal information. Moreover, attribute prompts further enhance the discriminative capacity of pedestrian features, leading to more robust and distinctive representations. Together, these innovations enable our model to achieve superior accuracy and generalization in person Re-identification tasks.

- **Result on MARS* Dataset.** We further evaluate our TriPro-ReID model on the MARS* dataset using both RGB and Event modalities. As a large and challenging benchmark for video-based person ReID, MARS* provides a rigorous testbed for assessing model performance. As shown in Table 3, our method achieves 88.4% mAP and 91.1% Rank-1 accuracy, surpassing all existing state-of-the-art approaches. These results clearly demonstrate the effectiveness of our framework in capturing discriminative spatiotemporal representations, significantly enhancing RGB-Event based person Re-ID performance.

5.4. Ablation Studies

- **Effect of Key Components.** To evaluate the effectiveness of each proposed component, we conduct ablation studies on the EvReID and MARS* datasets, as shown in Table 4. Starting from the base model (Index 1), we observe that introducing the Positive-Negative Attribute Prompt (PNAP) significantly improves performance on both datasets (Index 2). Notably, on EvReID, PNAP increases the mAP from 49.2% to 62.3% and Rank-1 from 73.0% to 81.1%. We attribute this remarkable gain to the degraded RGB modality in EvReID, which limits the effectiveness of traditional visual features. In such cases, injecting semantic attribute information greatly enriches the feature representation and compensates for the loss of visual cues. On MARS*, although the improvement brought by PNAP is more modest, it still demonstrates the general applicability of attribute prompts in well-illuminated settings. Adding the Cross Modal Prompts (CMP) alone (Index 3) also contributes positively compared to the base model, yielding consistent improvements on both datasets. CMP enhances the mAP by 1.0% on EvReID and 0.7% on MARS*, validating its role in facilitating RGB-Event interaction and capturing modality-complementary information. When both PNAP and CMP are integrated (Index 4), the full model achieves

Table 2. Comparison with public methods on our datasets. The highest results are shown in **bold**, while the second-best results are indicated with underline.

Methods	Modality	Publish	Code	mAP	Rank-1	Rank-5	Rank-10
#01 OSNet [58]	V+E	ICCV ₂₀₁₉	URL	23.7	49.1	65.4	72.3
#02 TCLNet [16]	V+E	ECCV ₂₀₂₀	URL	55.8	77.4	89.0	93.1
#03 AP3D [14]	V+E	ECCV ₂₀₂₀	URL	66.9	<u>86.5</u>	95.6	96.5
#04 MGH [53]	V+E	CVPR ₂₀₂₀	URL	43.2	70.9	89.4	92.7
#05 STMN [11]	V+E	ICCV ₂₀₂₁	URL	42.1	73.8	-	-
#06 PSTA [49]	V+E	ICCV ₂₀₂₁	URL	68.2	82.3	90.8	94.5
#07 GRL [29]	V+E	CVPR ₂₀₂₁	URL	38.9	62.6	78.0	83.6
#08 BiCnet-TKS [17]	V+E	CVPR ₂₀₂₁	URL	50.8	80.5	89.62	92.45
#09 SINet [6]	V+E	CVPR ₂₀₂₂	URL	50.2	77.4	92.8	96.2
#10 DCCT [30]	V+E	TNNLS ₂₀₂₃	URL	24.6	42.7	64.9	75.5
#11 CLIP-ReID [25]	V+E	AAAI ₂₀₂₃	URL	49.2	73.0	85.5	91.5
#12 SDCL [9]	V+E	CVPR ₂₀₂₃	URL	54.2	69.3	83.8	87.1
#13 TF-CLIP [54]	V+E	AAAI ₂₀₂₄	URL	56.9	78.6	91.8	94.3
#14 DeMo [50]	V+E	AAAI ₂₀₂₅	URL	59.4	75.7	90.1	92.8
#15 CLIMB-ReID [55]	V+E	AAAI ₂₀₂₅	URL	<u>68.3</u>	85.2	92.8	95.8
#16 TriPro-ReID (Ours)	V+E	-	URL	69.3	88.6	<u>94.3</u>	<u>95.4</u>

Table 3. Comparison with SOTA methods on MARS* dataset. The highest results are shown in **bold**, while the second-best results are indicated with underline.

Methods	Modality	mAP	Rank-1
#01 GRL [29]	V+E	82.8	88.7
#02 OSNet [58]	V+E	81.9	87.7
#03 SRS-Net [41]	V+E	83.8	89.3
#04 STMN [11]	V+E	83.4	89.0
#05 CTL [28]	V+E	85.3	89.6
#06 PSTA [49]	V+E	85.1	89.9
#07 SDCL [9]	V+E	<u>86.5</u>	<u>91.1</u>
#08 TriPro-ReID (Ours)	V+E	88.4	91.1

Table 4. Ablation study on our proposed EvReID dataset and the public MARS* dataset. PNAP and CMP denote **Positive and Negative Attribute Prompts** and **Cross Modal Prompts**, respectively, where the length of CMP is set as 20.

#Index	Base	PNAP	CMP	EvReID		MARS*	
				mAP	Rank-1	mAP	Rank-1
1	✓			49.2	73.0	86.8	88.7
2	✓	✓		62.3	81.1	87.2	89.9
3	✓		✓	50.2	75.2	87.5	90.1
4	✓	✓	✓	69.3	88.6	88.4	91.1

the best performance. On EvReID, mAP and Rank-1 reach 69.3% and 88.6%, respectively, while on MARS*, the model attains 88.4% mAP and 91.1% Rank-1. These results demonstrate that PNAP and CMP are complementary modules, and their combination effectively boosts representation learning in both modality-degraded and modality-rich scenarios.

Table 5. Comparing different Attribute Prompt settings on EvReID dataset.

Settings	mAP	Rank-1	Rank-5	Rank-10
PNAP	69.3	88.6	94.3	95.4
Only Positive	54.4	78.9	89.0	91.5
w/o Attribute Context Prompts	51.4	77.4	89.6	91.2

- **Effect of Different Attribute Prompt Settings.** Table 5 presents a comparative analysis of different attribute prompt configurations on the EvReID dataset. The complete PNAP setting, which includes both positive and negative attribute prompts with contextual information, achieves the best performance, reaching 69.3% mAP and 88.6% Rank-1 accuracy. This confirms the effectiveness of bidirectional semantic supervision in enhancing identity discrimination. When using only positive attribute prompts, the performance drops to 54.4% mAP and 78.9% Rank-1. This is mainly because some attributes may be shared by multiple individuals, potentially causing identity confusion. The addition of negative prompts further refines semantic boundaries and helps suppress irrelevant or misleading features, resulting in more robust representation learning. Furthermore, removing attribute context prompts entirely leads to a further drop in performance (51.1% mAP and 77.4% Rank-1). This suggests that the contextual tokens play a crucial role in enhancing the expressiveness of attribute semantics. Without them, the prompts become isolated keywords lacking sufficient semantic richness, which weakens the alignment between textual descriptions and visual representations. Therefore, context prompts are essential for providing richer and more discriminative semantic guidance.

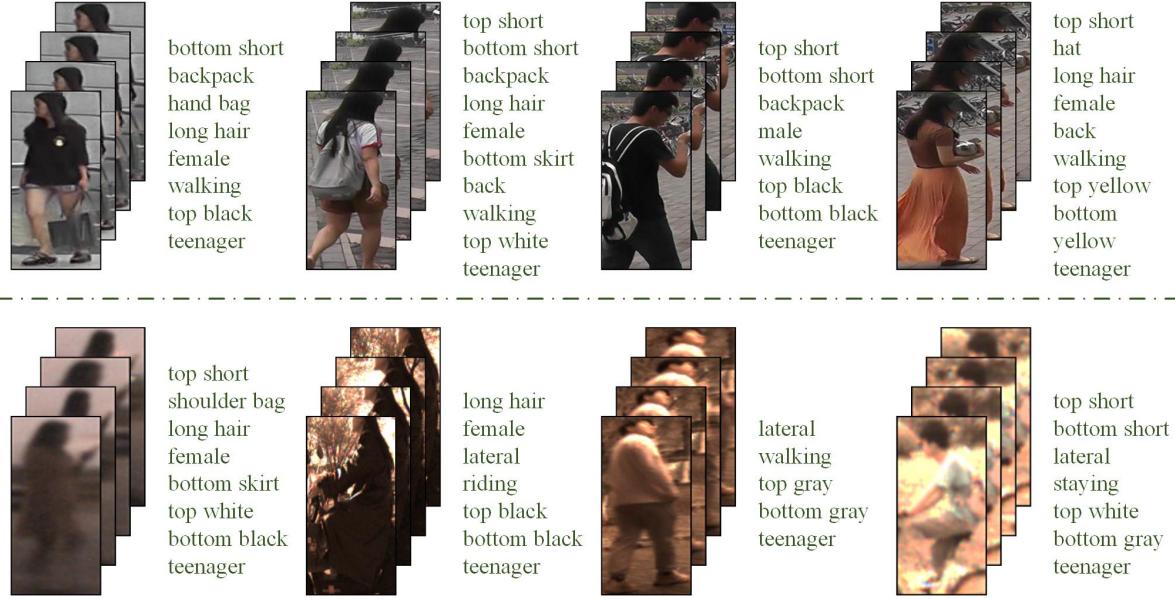


Figure 4. Visualization of pedestrian attribute prediction results. The first row shows the predictions from the MARS dataset, while the second row displays the predictions from the EvReID dataset.

Table 6. Comparing different Depths and Lengths of Cross-modal Prompt on EvReID dataset.

Length	mAP	Rank-1	Rank-5	Rank-10
40	62.4	81.4	90.9	93.9
20	69.3	88.6	94.3	95.4
10	69.8	87.1	92.8	95.4
Depth	mAP	Rank-1	Rank-5	Rank-10
12	69.3	88.6	94.3	95.4
6	66.7	85.8	93.4	95.3
3	66.2	84.9	93.7	95.3

• **Effect of Different Depth and Length of CMP.** We investigate the influence of the Cross Modal Prompt (CMP)’s length and depth on the EvReID dataset. As shown in Table 6, varying these configurations leads to distinct changes in performance. For prompt length, we observe that shorter lengths yield better results. Reducing the CMP length from 40 to 20 leads to improvement in mAP (from 62.4% to 69.3%) and Rank-1 accuracy (from 81.4% to 88.6%). Further shortening the length to 10 achieves the best performance, with 69.8% mAP and 87.1% Rank-1. This trend suggests that overly long prompts may introduce redundant or noisy information, while compact prompts encourage more focused and effective cross-modal interactions. For prompt depth, while a depth of 12 yields strong results (69.3% mAP, 88.6% Rank-1), reducing it to 6 and 3 causes nearly 3 score drops in accuracy.

• **Effect of Different Projectors for CMP.** We study the impact of different projection strategies in the design of

Table 7. Comparing different Projectors of Cross-Modal Prompt on EvReID dataset, where the length of CMP is set as 20.

Settings	mAP	Rank-1	Rank-5	Rank-10
FC	69.3	88.6	94.3	95.4
Adapters	69.7	89.7	94.3	94.3
None	68.8	86.3	94.7	95.1

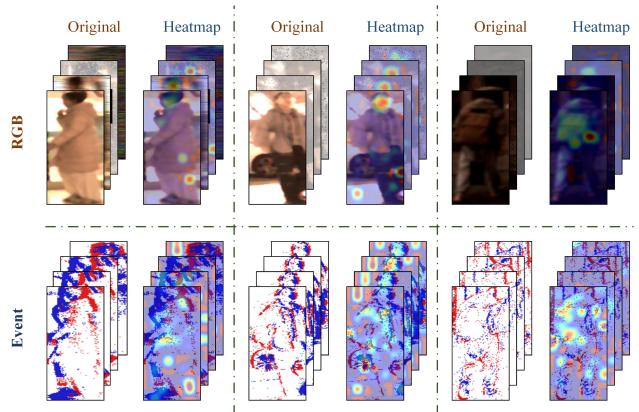


Figure 5. Visualization of pedestrian heatmap results. The first row shows the heatmap for the RGB modality, while the second row displays the corresponding heatmap for the Event modality.

Cross-Modal Prompts by implementing three variants, including using a fully connected (FC) layer, an adapter module, and a direct mapping approach without projection. As reported in Table 7, the adapter-based strategy achieves the

Table 8. Comparing different Cross Modal Type settings on MARS* dataset.

Settings	mAP	Rank-1
RGB → Event	88.4	91.1
Event → RGB	87.9	91.1
Bi-direction	87.4	90.3

best performance, with a mAP of 69.7% and Rank-1 accuracy of 89.7%, outperforming both the FC-based (69.3% / 88.6%) and the non-projection setting (68.8% / 86.3%).

- **Effect of Different CMP Settings.** To evaluate the influence of different interaction directions in the Cross-Modal Prompt (CMP) module, we compare three representative configurations: unidirectional interaction from RGB to event modality, unidirectional interaction from event to RGB, and bidirectional interaction. Experimental results on the MARS* dataset, presented in Table 8, indicate that guiding the event branch using prompts generated from the RGB modality achieves the best performance, with 88.4% mAP and 91.1% Rank-1 accuracy.

5.5. Visualization

- **Results of Attributes Prediction.** As shown in Fig. 4, we visualize the attribute prediction results on MARS and EvReID datasets. Although the predictions are not entirely comprehensive or perfectly accurate, the pre-trained attribute recognition model is able to capture most of the salient pedestrian attributes. These include clothing types, accessories, and coarse physical characteristics. Such coverage provides sufficiently rich semantic cues, which serve as effective prompts for identity representation learning in our framework.

- **Discriminative Attention Maps.** In Fig. 5, we visualize the learned feature maps to highlight how our network effectively focuses on discriminative, fine-grained details of pedestrians. The attention maps emphasize critical body parts, which are essential for accurate person re-identification, especially under occlusions or pose variations. Our dual-modality approach, combining RGB and event stream data, further enhances this fine-grained focus. RGB provides rich appearance features, while the event stream captures motion and temporal changes. By jointly attending to both modalities, the network integrates complementary information, improving feature representation and alignment between appearance and motion, which boosts performance.

- **Visualization of Rank List.** Fig. 6 shows the rank lists from TriPro-ReID, demonstrating its ability to produce accurate rankings and validate its effectiveness. We present the top-5 ranked results for each query, highlighting the model’s precision in ranking and its effectiveness in identifying the correct matches.

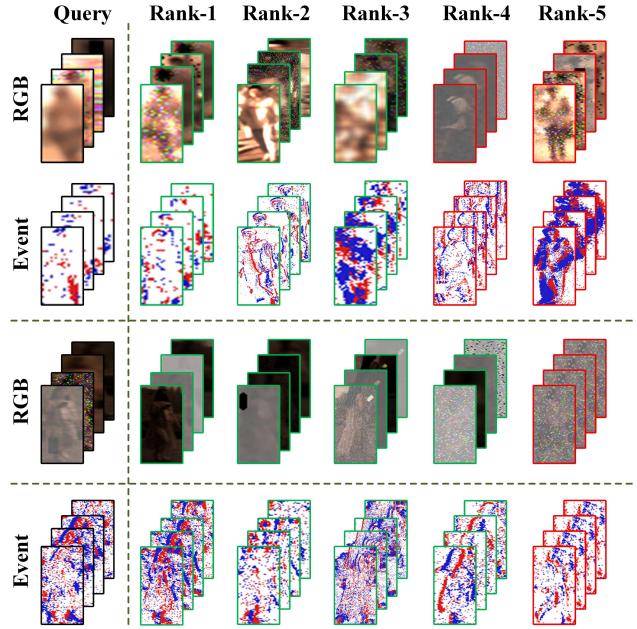


Figure 6. Visualization of the rank list of person re-identification.

5.6. Limitation Analysis

Although the proposed TriPro-ReID achieves promising results on RGB-Event person re-identification, it still has some limitations. First, the Positive-Negative Attribute Prompt (PNAP) relies on a pre-trained attribute recognition model, whose accuracy and generalization may be limited. Incorrect predictions can introduce noise into the prompts and affect feature learning. Moreover, attribute labels are often coarse and shared among different identities, which reduces the discriminative ability of the prompts. Second, the three-stage training pipeline increases overall complexity. Each stage involves specific objectives and model freezing or unfreezing, requiring careful tuning and longer training time. Finally, the use of large vision-language models like CLIP leads to high memory and computation costs, which may hinder deployment on resource-limited platforms.

6. Conclusion and Future Works

This paper presents TriPro-ReID, a novel approach for RGB-Event-based person re-identification, which integrates an attribute-guided framework. Our method demonstrates superior performance, addressing key challenges in existing dual-modality person ReID. Furthermore, we introduce a new dataset, EvReID, designed to foster progress in this domain by providing a comprehensive benchmark that spans diverse seasons, scenes, and lighting conditions. In addition, we retrain 15 state-of-the-art methods on this dataset, contributing to the ongoing development of RGB-Event-

based person re-identification. In the future, we plan to expand the scale of the dataset and investigate the impact of temporal dynamics on model performance. Additionally, the use of more advanced large foundation models [45] will be explored to enhance the overall accuracy.

References

- [1] Shafiq Ahmad, Gianluca Scarpellini, Pietro Morerio, and Alessio Del Bue. Event-driven re-id: A new benchmark and method towards privacy-preserving person re-identification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 459–468, 2022.
- [2] Shafiq Ahmad, Pietro Morerio, and Alessio Del Bue. Person re-identification without identification via event anonymization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11132–11141, 2023.
- [3] Shafiq Ahmad, Pietro Morerio, and Alessio Del Bue. Person re-identification without identification via event anonymization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11132–11141, 2023.
- [4] Shafiq Ahmad, Pietro Morerio, and Alessio Del Bue. Event anonymization: privacy-preserving person re-identification and pose estimation in event-based vision. *IEEE Access*, 2024.
- [5] Abhishek Aich, Meng Zheng, Srikrishna Karanam, Terrence Chen, Amit K Roy-Chowdhury, and Ziyang Wu. Spatio-temporal representation factorization for video-based person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 152–162, 2021.
- [6] Shutao Bai, Bingpeng Ma, Hong Chang, Rui Huang, and Xilin Chen. Salient-to-broad transition for video person re-identification. In *CVPR*, 2022.
- [7] Shutao Bai, Bingpeng Ma, Hong Chang, Rui Huang, and Xilin Chen. Salient-to-broad transition for video person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7339–7348, 2022.
- [8] Alina Bialkowski, Simon Denman, Sridha Sridharan, Clinton Fookes, and Patrick Lucey. A database for person re-identification in multi-camera surveillance networks. In *2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2012.
- [9] Chengzhi Cao, Xueyang Fu, Hongjian Liu, Yukun Huang, Kunyu Wang, Jiebo Luo, and Zheng-Jun Zha. Event-guided person re-identification via sparse-dense complementary learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17990–17999, 2023.
- [10] Di Chen, Andreas Doering, Shanshan Zhang, Jian Yang, Juergen Gall, and Bernt Schiele. Keypoint message passing for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 239–247, 2022.
- [11] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsuk Ham. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12036–12045, 2021.
- [12] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.
- [13] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13884–13893, 2023.
- [14] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *ECCV*, 2020.
- [15] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis: 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, May 2011. Proceedings 17*, pages 91–102. Springer, 2011.
- [16] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *ECCV*, 2020.
- [17] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *CVPR*, 2021.
- [18] Xiang Hu, Yuhao Wang, Pingping Zhang, and Huchuan Lu. Latex: Leveraging attribute-based text knowledge for aerial-ground person re-identification. *arXiv preprint arXiv:2503.23722*, 2025.
- [19] Ju Huang, Shiao Wang, Shuai Wang, Zhe Wu, Xiao Wang, and Bo Jiang. Mamba-fetrack: Frame-event tracking via state space model. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 3–18. Springer, 2024.
- [20] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022.
- [21] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023.
- [22] Seijo Kim, Seongsik Park, Byunggook Na, and Sungroh Yoon. Spiking-yolo: spiking neural network for energy-efficient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11270–11277, 2020.
- [23] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3958–3967, 2019.
- [24] Renkai Li, Xin Yuan, Wei Liu, and Xin Xu. Event-based video person re-identification via cross-modality and temporal collaboration. *arXiv preprint arXiv:2501.07296*, 2025.

- [25] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels. *arXiv preprint arXiv:2211.13977*, 2022.
- [26] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1405–1413, 2023.
- [27] Jiawei Liu, Zheng-Jun Zha, Wei Wu, Kecheng Zheng, and Qibin Sun. Spatial-temporal correlation and topology learning for person re-identification in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4370–4379, 2021.
- [28] Jiawei Liu, Zheng-Jun Zha, Wei Wu, Kecheng Zheng, and Qibin Sun. Spatial-temporal correlation and topology learning for person re-identification in videos, 2021.
- [29] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, and Xiaoyun Yang. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13334–13343, 2021.
- [30] Xuehu Liu, Chenyang Yu, Pingping Zhang, and Huchuan Lu. Deeply coupled convolution-transformer with spatial-temporal complementary learning for video-based person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [32] Xinhao Luo, Man Yao, Yuhong Chou, Bo Xu, and Guoqi Li. Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection. In *European Conference on Computer Vision*, pages 253–272. Springer, 2024.
- [33] Athira Nambiar, Matteo Taiana, Dario Figueira, Jacinto C Nascimento, and Alexandre Bernardino. A multi-camera video dataset for research on high-definition surveillance. *International Journal of Machine Intelligence and Sensory Signal Processing*, 1(3):267–286, 2014.
- [34] Yansong Peng, Yueyi Zhang, Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Get: Group event transformer for event-based vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6038–6048, 2023.
- [35] Yansong Peng, Hebei Li, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. Scene adaptive sparse transformer for event-based object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16794–16804, 2024.
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [37] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016.
- [38] Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4390–4403, 2021.
- [39] Qiaoyi Su, Yuhong Chou, Yifan Hu, Jianing Li, Shijie Mei, Ziyang Zhang, and Guoqi Li. Deep directly-trained spiking neural networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6555–6565, 2023.
- [40] Bin Wang, Yuying Liang, Lei Cai, Huakun Huang, and Huanqiang Zeng. Image re-identification: Where self-supervision meets vision-language learning. *Image and Vision Computing*, 154:105415, 2025.
- [41] Haoran Wang, Licheng Jiao, Shuyuan Yang, Lingling Li, and Zexin Wang. Simple and effective: Spatial rescaling for person reidentification. *IEEE Transactions on neural networks and learning systems*, 33(1):145–156, 2020.
- [42] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part IV 13*, pages 688–703. Springer, 2014.
- [43] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2540–2549, 2022.
- [44] Xiao Wang, Shaofei Zheng, Rui Yang, Aihua Zheng, Zhe Chen, Jin Tang, and Bin Luo. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121:108220, 2022.
- [45] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482, 2023.
- [46] Xiao Wang, Ju Huang, Shiao Wang, Chuanming Tang, Bo Jiang, Yonghong Tian, Jin Tang, and Bin Luo. Long-term frame-event visual tracking: Benchmark dataset and baseline. *arXiv preprint arXiv:2403.05839*, 2024.
- [47] Xiao Wang, Yu Jin, Wentao Wu, Wei Zhang, Lin Zhu, Bo Jiang, and Yonghong Tian. Object detection using event camera: A moe heat conduction based detector and a new benchmark dataset, 2024.
- [48] Xiao Wang, Qian Zhu, Jiandong Jin, Jun Zhu, Futian Wang, Bo Jiang, Yaowei Wang, and Yonghong Tian. Spatio-temporal side tuning pre-trained foundation models for video-based pedestrian attribute recognition. *arXiv preprint arXiv:2404.17929*, 2024.
- [49] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid spatial-temporal aggregation for video-based person re-identification. In *ICCV*, 2021.
- [50] Yuhao Wang, Yang Liu, Aihua Zheng, and Pingping Zhang. Decoupled feature-based mixture of experts for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8141–8149, 2025.

- [51] Yuhao Wang, Yongfeng Lv, Pingping Zhang, and Huchuan Lu. Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [52] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5177–5186, 2018.
- [53] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 2896–2905. IEEE, 2020.
- [54] Chenyang Yu, Xuehu Liu, Yingquan Wang, Pingping Zhang, and Huchuan Lu. Tf-clip: Learning text-free clip for video-based person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6764–6772, 2024.
- [55] Chenyang Yu, Xuehu Liu, Jiawen Zhu, Yuhao Wang, Pingping Zhang, and Huchuan Lu. Climb-reid: A hybrid clip-mamba framework for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9589–9597, 2025.
- [56] Quan Zhang, Lei Wang, Vishal M Patel, Xiaohua Xie, and Jianhaung Lai. View-decoupled transformer for person re-identification under aerial-ground camera network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22000–22009, 2024.
- [57] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VI 14*, pages 868–884. Springer, 2016.
- [58] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification, 2019.
- [59] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.
- [60] Yabin Zhu, Xiao Wang, Chenglong Li, Bo Jiang, Lin Zhu, Zhixiang Huang, Yonghong Tian, and Jin Tang. Crsot: Cross-resolution object tracking using unaligned frame and event cameras. *arXiv preprint arXiv:2401.02826*, 2024.