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

Xiao Wang[1], Qian Zhu[1], Shujuan Wu[1], Bo Jiang[1]*, Shiliang Zhang[2, 3]

[1]School of Computer Science and Technology, Anhui University, Hefei, China, 230601
[2]Peng Cheng Laboratory, Shenzhen, China, 499100
[3]School of Computer Science, Peking University, China, 100871
xiaowang@ahu.edu.cn, zq542664@163.com, wushujuan0114@163.com, jiangbo@ahu.edu.cn, slzhang.jdl@pku.edu.cn

Abstract

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.

Code — https://github.com/Event-AHU/Neuromorphic_ReID/tree/main/TriPro-main

Datasets — https://github.com/Event-AHU/Neuromorphic_ReID/tree/main/TriPro-main

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.

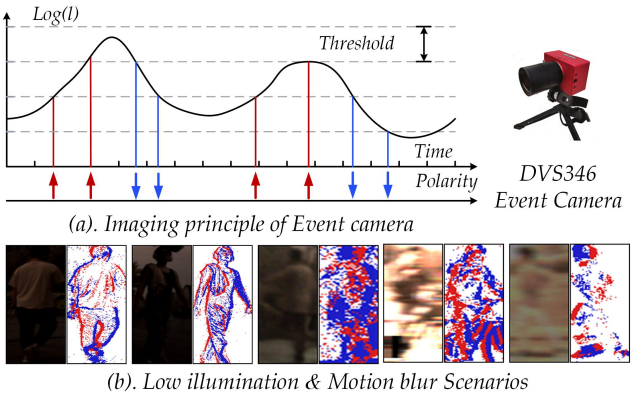

Figure 1: Imaging principle of event camera and comparison of RGB frames and event streams in challenging scenarios.

To address the aforementioned issues, some researchers resort to the bio-inspired event cameras for person re-identification (Xu et al. 2025; Deng et al. 2025c,b), due to their advantages in low energy consumption, high dynamic range, no motion blur, and spatial sparsity, as shown in Fig. 1. Specifically, Ahmad et al. (Ahmad, Morerio, and Del Bue 2023) 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. (Cao et al. 2023) 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. (Li et al. 2025) 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 issues: 1). Existing person ReID algorithms are typically trained and evaluated on small-scale or simulated event cam-

*Corresponding Author: Bo Jiang & Shiliang Zhang

era datasets, making it difficult to assess their real identification performance and generalization ability. 2). Current ReID algorithms mainly focus on learning event stream features (Li et al. 2025) or fusing RGB and event features (Cao et al. 2023), but fail to consider the semantic information, such as pedestrian attributes *long hair, wearing glasses* (Wang et al. 2022; Wu et al. 2024, 2025b), 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 (Ahmad, Morerio, and Del Bue 2023) 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) 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++ (Wang et al. 2024b) 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.

Related Works

Event-based Vision

While RGB-based vision systems generally achieve strong performance, they often fail in low-light conditions. To address this, event-based methods (Cheng, Knoll, and Cao 2025) have emerged, leveraging the high temporal resolution and low latency of event data. In person ReID, Ahmad et al. (Ahmad, Morerio, and Del Bue 2023) introduce the first event-based dataset and propose a privacy-preserving event ReID framework. In detection and tracking, recent works such as RVT (Gehrig and Scaramuzza 2023), MvHeat-DET (Wang et al. 2024a), and Mamba-FETrack (Huang et al. 2024) demonstrate the potential of event data through advanced architectures like Transformers, SNNs, and Mamba networks. Despite their advantages, event-based approaches often underperform in accuracy compared to RGB methods due to limited semantic content. To bridge this gap, we propose a dual-branch RGB-Event framework that exploits the complementary strengths of both modalities for enhanced performance.

Person Re-Identification

Person re-identification (ReID) (Shu et al. 2021; Zheng et al. 2023; Deng et al. 2025a) has gained attention, especially in video-based settings leveraging temporal coherence. Early CNN-based methods like AP3D (Gu et al. 2020) and TCLNet (Hou et al. 2020) model spatial-temporal cues, while CTL (Liu et al. 2021a) and PSTA (Wang et al. 2021) incorporate LSTMs for temporal modeling. Graph-based approaches such as MGH (Yan et al. 2020) and keypoint-based GCNs (Chen et al. 2022) capture complex spatial-temporal relations. Transformer-based models like DCCT (Liu et al. 2023) and VDT (Zhang et al. 2024) enhance global context representation. Event-based ReID methods (Ahmad, Morerio, and Del Bue 2023) leverage event streams for accuracy under challenging conditions. Vision-language models like CLIP (Radford et al. 2021) have been adapted for ReID with textual supervision (Yu et al. 2024), and semantic knowledge integration (Wang et al. 2025b). Unlike heavy semantic models, we propose a lightweight attribute-guided dual-modality framework that predicts semantic attributes, encodes them as text features, and fuses them with visual cues via cross-modal prompting for fine-grained, semantically enriched representations.

#Index	Dataset	Year	#IDs	#Images	Modality	Seasons	Lighting	Real Event
01	**PRID-2011** (Hirzer et al. 2011)	2011	200	40,033	RGB	Single	Daytime	✗
02	**SAIVT-Softbio** (Bialkowski et al. 2012)	2012	152	64,472	RGB	Single	Daytime	✗
03	**iLIDS-VID** (Wang et al. 2014)	2014	300	42,460	RGB	Single	Daytime	✗
04	**HDA Person** (Nambiar et al. 2014)	2014	53	2,976	RGB	Single	Daytime	✗
05	**MARS** (Zheng et al. 2016)	2016	1,261	1,067,516	RGB	Single	Daytime	✗
06	**DukeMTMC-VideoReID** (Wu et al. 2018)	2018	1,404	815,420	RGB	Single	Daytime	✗
07	**LS-VID** (Li et al. 2019)	2019	3,772	2,982,685	RGB	Single	**Day, Night**	✗
08	**Event ReID** (Ahmad, Morerio, and Del Bue 2023)	2023	33	16,000	Event	Single	Daytime	✗
09	**EvReID (Ours)**	2025	1,200	118,988	**RGB-Event**	**Multi**	**Day, Night**	✓

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.

Benchmark Datasets for Video-Based Person Re-Identification

The most commonly used datasets for video-based person re-identification (ReID) include PRID-2011 (Hirzer et al. 2011), iLIDS-VID (Wang et al. 2014), MARS (Zheng et al. 2016), DukeMTMC-VideoReID (Wu et al. 2018), and LS-VID (Li et al. 2019). PRID-2011 and iLIDS-VID contain a few hundred sequences captured by two cameras under controlled and surveillance conditions, respectively. MARS is a large-scale benchmark with over 17,000 tracklets of 1,261 identities from six cameras. DukeMTMC-VideoReID offers 702 training and 702 testing identities with additional distractors. LS-VID further expands the scale with 3,772 annotated identities in diverse scenarios. Beyond RGB data, event modalities have been introduced, with Ahmad et al. (Ahmad, Morerio, and Del Bue 2023) proposing the first event-based ReID dataset, Event-ReID, containing event sequences for 33 individuals. Existing datasets are limited either by modality or scale, and none combine RGB and event data under a unified setting. To fill this gap, we present EvReID, a dual-modality video ReID benchmark supporting RGB, event, and their fusion, enabling comprehensive evaluation and advancing multi-modal person ReID research.

EvReID Benchmark Dataset

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.

Data Collection and Statistical Analysis

s As shown in Table 1, we compare several representative video-based person re-identification (ReID) datasets with our proposed EvReID dataset. The details about our proposed EvReID dataset can be found in the supplementary materials.

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, and more details can be found in the supplementary materials.

Methodology

Preliminary: CLIP and CLIP-ReID

CLIP (Radford et al. 2021) consists of a visual encoder and a text encoder trained jointly with contrastive learning to map images and texts into a shared embedding space. Given a batch of paired samples, images and texts are encoded and projected into unified feature embeddings. The similarity between image and text features is computed by their projected embeddings. Contrastive losses align these features during training. In downstream tasks, text inputs are usually prompts like "A photo/video of a {class}," where the class is specified. However, in person ReID, identities are represented as integer labels rather than descriptive text. To address this, CLIP-ReID (Li, Sun, and Li 2023) introduces learnable identity-specific tokens forming personalized prompts such as "A photo of a $[X]_1, [X]_2, \cdots, [X]_{Num}$ person." The encoders are frozen while optimizing these tokens to learn discriminative textual identity representations.

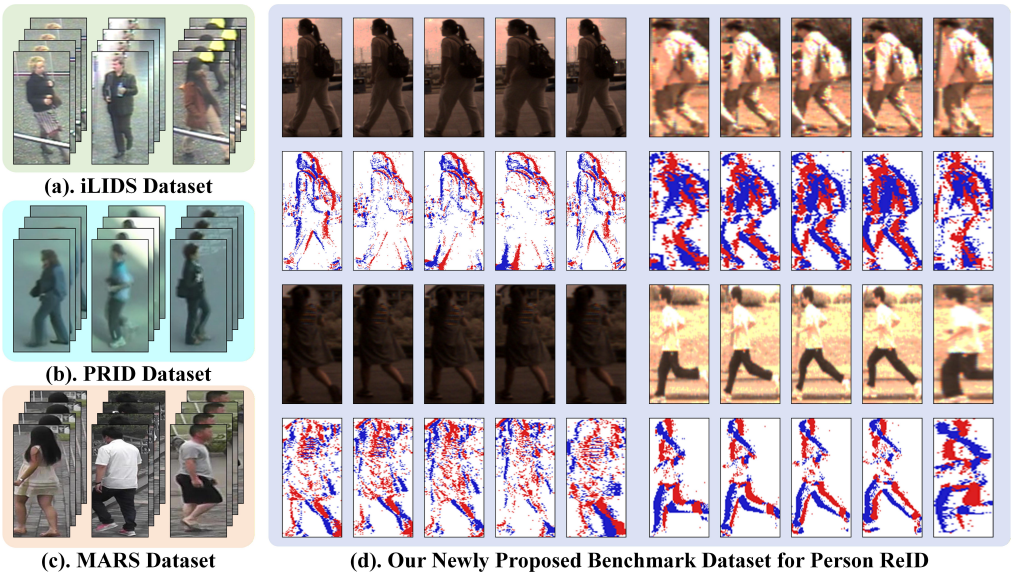

Figure 2: Comparison of existing video-based ReID datasets and our newly proposed EvReID dataset.

Although effective for RGB-based ReID, a simple dual-branch CLIP-ReID architecture is insufficient for RGB-Event dual-modality scenarios, as it cannot fully exploit complementary temporal and modality-specific cues from event data. Moreover, these identity-aware prompts do not explicitly capture fine-grained pedestrian attribute information available in video frames.

Overview

We propose a novel framework, **Tri**ple-**Pro**mpting **Re-id**entification (TriPro-ReID), comprising two core modules: Positive-Negative Attribute Prompting (PNAP) and Cross-Modal Prompting (CMP). The framework is trained in three stages. First, we use the vanilla CLIP contrastive learning pipeline to initialize identity-aware context prompts for each person and associate them with corresponding IDs to obtain textual representations. Second, we fix the identity prompts and introduce visual prompts into the visual encoder to align visual features with textual ones. Meanwhile, CMPs bind RGB and Event visual features with identity embeddings, enabling cross-modality prompt propagation to leverage complementary multimodal information and extract generalized knowledge from the pre-trained CLIP model. Third, PNAPs, generated from person attributes and encoded by the text encoder, are injected into intermediate visual encoder layers to dynamically enhance visual features with attribute-aware discrimination. The model is supervised using ID and triplet losses for person ReID. Detailed descriptions of these modules follow.

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, we aim to extract identity-specific sequence representations. For an identity y_i, we use the pre-trained CLIP visual encoder to extract sequence-level features from RGB and Event video sequences, denoted as $\hat{F}_R$ and $\hat{F}_E$. For RGB, each image sequence $\{I_{R_t}\}_{t=1}^{T}$ is divided into N_p non-overlapping patches. Each patch is projected into token embeddings via a linear projection layer, and a learnable class token [CLS] is added to form the input to the visual encoder:

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

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:

$$\hat{F}_R = \mathcal{V}(Z_{R_t}^{\text{Patch}}) \quad (2)$$

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 $[X]_1, [X]_2, \cdots, [X]_{\text{Num}}$ person", where each $[X]_{y_i}$ ($y_i \in \{1, \ldots, \text{Num}\}$) 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:

$$T_{y_i} = \mathcal{T}(\text{text}_{y_i}) \quad (3)$$

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

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

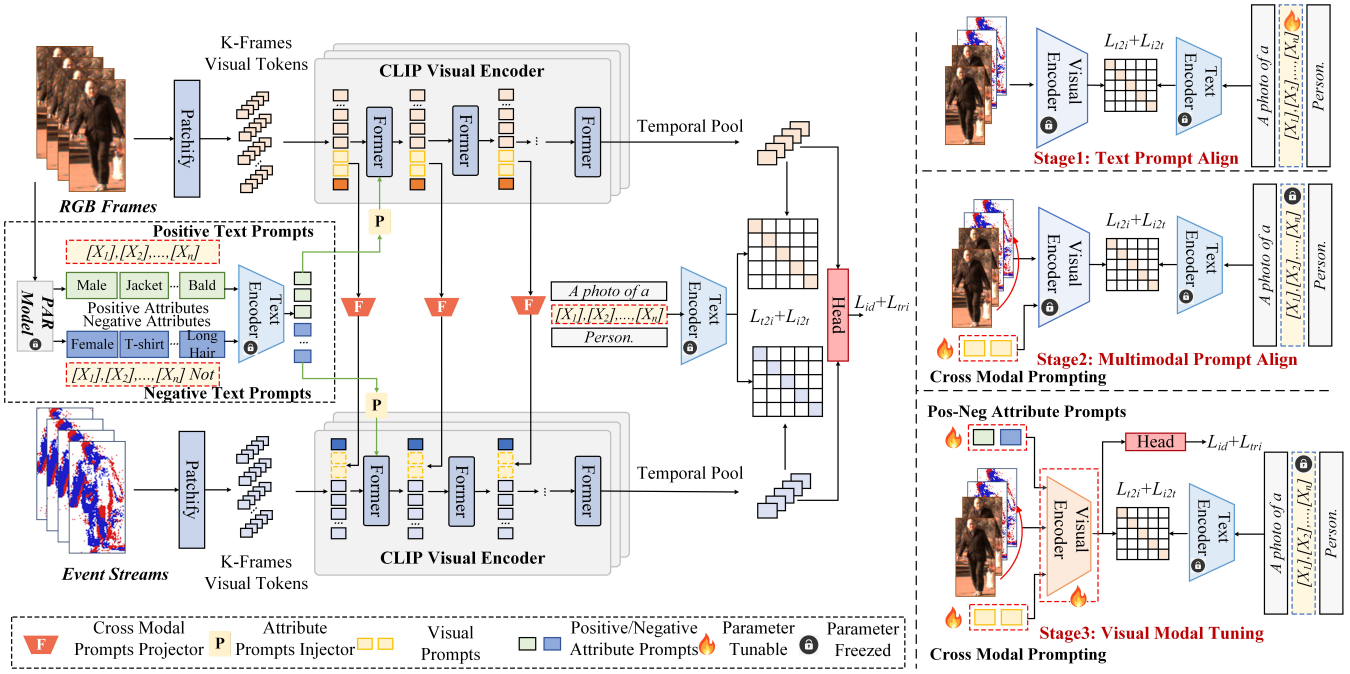

Figure 3: An illustration of our proposed TriPro-ReID framework. The left part illustrates the overall architecture of the RGB-Event person re-identification framework, consisting of dual-stream backbones and multi-level fusion modules. The right part presents the three-stage training strategy, including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

with fine-grained semantic cues, we propose a *Positive-Negative Attribute Prompting* (PNAP) mechanism, which is inspired by the learnable context prompt strategy in (Zhou et al. 2022; Khattak et al. 2023; Wu et al. 2025a), and is further motivated by (Wang et al. 2025b), 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 (Wang et al. 2024b), 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 (Zhou et al. 2022), 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 $\hat{F}_R$ and $\hat{F}_E$ to yield prompt-enhanced identity representations F_R^0 and F_E^0. The overall process can be summarized as:

$$PNAP = \mathcal{T}(\mathcal{A}^{\text{pos}}, P^{\text{pos}}) \cup \mathcal{T}(\mathcal{A}^{\text{neg}}, P^{\text{neg}}) \tag{4}$$

$$F_R^0 = \text{Con}(\hat{F}_R, \text{FC}(PNAP)) \tag{5}$$

$$F_E^0 = \text{Con}(\hat{F}_E, \text{FC}(PNAP)) \tag{6}$$

Here, $\mathcal{A} = \mathcal{PAR}(I_R)$ is the recognition result from the pedestrian 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 (Jia et al. 2022), 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-injected features F_R^1 and F_E^1, respectively:

$$F_R^1 = \text{Con}(F_R^0, CMP_R^0) \tag{7}$$

$$CMP_E^0 = \text{FC}(CMP_R^0) \tag{8}$$

$$F_E^1 = \text{Con}(F_E^0, CMP_E^0) \tag{9}$$

Here, F_R^0 and F_E^0 denote the visual features at the input 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 CMPs are iteratively updated to $CMP_R^1, \ldots, CMP_R^k$, where the

Methods	Modality	Publish	mAP	Rank-1	Rank-5	Rank-10
#01 AP3D (Gu et al. 2020)	V	$ECCV_{2020}$	65.4	83.0	92.1	95.3
#02 PSTA (Wang et al. 2021)	V	$ICCV_{2021}$	63.3	85.2	92.5	95.3
#03 GRL (Liu et al. 2021b)	V	$CVPR_{2021}$	34.5	58.2	73.6	84.0
#04 SINet (Bai et al. 2022)	V	$CVPR_{2022}$	62.7	83.0	92.4	**96.5**
#05 CLIP-ReID (Li, Sun, and Li 2023)	V	$AAAI_{2023}$	49.9	68.8	81.0	84.0
#01 AP3D (Gu et al. 2020)	E	$ECCV_{2020}$	40.6	67.3	80.8	86.2
#02 PSTA (Wang et al. 2021)	E	$ICCV_{2021}$	37.1	61.0	77.4	84.6
#03 GRL (Liu et al. 2021b)	E	$CVPR_{2021}$	38.9	62.6	78.0	83.6
#04 SINet (Bai et al. 2022)	E	$CVPR_{2022}$	40.1	67.0	81.5	85.2
#05 CLIP-ReID (Li, Sun, and Li 2023)	E	$AAAI_{2023}$	30.4	52.5	70.3	79.1
#01 OSNet (Zhou et al. 2019)	V+E	$ICCV_{2019}$	23.7	49.1	65.4	72.3
#02 TCLNet (Hou et al. 2020)	V+E	$ECCV_{2020}$	55.8	77.4	89.0	93.1
#03 AP3D (Gu et al. 2020)	V+E	$ECCV_{2020}$	66.9	<u>86.5</u>	**95.6**	**96.5**
#04 MGH (Yan et al. 2020)	V+E	$CVPR_{2020}$	43.2	70.9	89.4	92.7
#05 STMN (Eom et al. 2021)	V+E	$ICCV_{2021}$	42.1	73.8	-	-
#06 PSTA (Wang et al. 2021)	V+E	$ICCV_{2021}$	68.2	82.3	90.8	94.5
#07 GRL (Liu et al. 2021b)	V+E	$CVPR_{2021}$	38.9	62.6	78.0	83.6
#08 BiCnet-TKS (Hou et al. 2021)	V+E	$CVPR_{2021}$	50.8	80.5	89.6	92.5
#09 SINet (Bai et al. 2022)	V+E	$CVPR_{2022}$	67.1	83.4	93.0	95.6
#10 DCCT (Liu et al. 2023)	V+E	$TNNLS_{2023}$	24.6	42.7	64.9	75.5
#11 CLIP-ReID (Li, Sun, and Li 2023)	V+E	$AAAI_{2023}$	49.2	73.0	85.5	91.5
#12 SDCL (Cao et al. 2023)	V+E	$CVPR_{2023}$	54.2	69.3	83.8	87.1
#13 TF-CLIP (Yu et al. 2024)	V+E	$AAAI_{2024}$	56.9	78.6	91.8	94.3
#14 DeMo (Wang et al. 2025a)	V+E	$AAAI_{2025}$	59.4	75.7	90.1	92.8
#15 CLIMB-ReID (Yu et al. 2025)	V+E	$AAAI_{2025}$	<u>68.3</u>	85.2	92.8	<u>95.8</u>
#16 TriPro-ReID (Ours)	V+E	-	**69.3**	**88.6**	<u>94.3</u>	95.4

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

superscript k indicates the corresponding transformer layer. At each layer, the CMPs in the Event branch are synchronously replaced with the transformed RGB CMPs from the same layer, enabling continuous and guided feature fusion between the two modalities throughout the network.

Experiments

Datasets and Evaluation Metric

To evaluate the performance, we conduct a comprehensive benchmark of 15 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. (Cao et al. 2023) generate events from MARS (Zheng et al. 2016). 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-$\mathcal{K}$ ($\mathcal{K} = 1, 5, 10$) as our evaluation metrics.

Implementation Details

We adopt CLIP-ReID as our baseline, using separate backbone parameters for each modality. We use SGD in the first two stages and AdamW (Loshchilov and Hutter 2017) in the third, with learning rates of 3.5×10^{-3}, 3.5×10^{-3}, and 5×10^{-6}, and batch sizes of 64, 64, and 8, respectively. All experiments are conducted on an NVIDIA RTX 3090 GPU. Additional details are provided in the source code and supplementary materials.

Methods	Modality	mAP	Rank-1
#01 OSNet (Zhou et al. 2019)	V	81.4	87.3
#02 GRL (Liu et al. 2021b)	V	84.8	91.0
#03 STMN (Eom et al. 2021)	V	84.5	90.5
#04 PSTA (Wang et al. 2021)	V	85.8	91.5
#05 TF-CLIP (Yu et al. 2024)	V	89.4	93.0
#06 CLIMB-ReID (Yu et al. 2025)	V	89.7	93.3
#07 TCLNet (Hou et al. 2020)	E	38.2	25.3
#08 GRL (Liu et al. 2021b)	E	27.7	16.7
#09 BiCnet-TKS (Hou et al. 2021)	E	30.9	17.3
#10 STMN (Eom et al. 2021)	E	22.4	10.0
#11 OSNet (Zhou et al. 2019)	V+E	81.9	87.7
#12 GRL (Liu et al. 2021b)	V+E	82.8	88.7
#13 SRS-Net (Wang et al. 2020)	V+E	83.8	89.3
#14 STMN (Eom et al. 2021)	V+E	83.4	89.0
#15 PSTA (Wang et al. 2021)	V+E	85.1	89.9
#16 SDCL (Cao et al. 2023)	V+E	86.5	91.1
#17 TriPro-ReID (Ours)	V+E	88.4	91.1

Table 3: Comparison on the MARS* dataset.

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 complemen-

#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**

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

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

Table 5: Comparing different Depths and Lengths of Cross-modal Prompt on EvReID dataset.

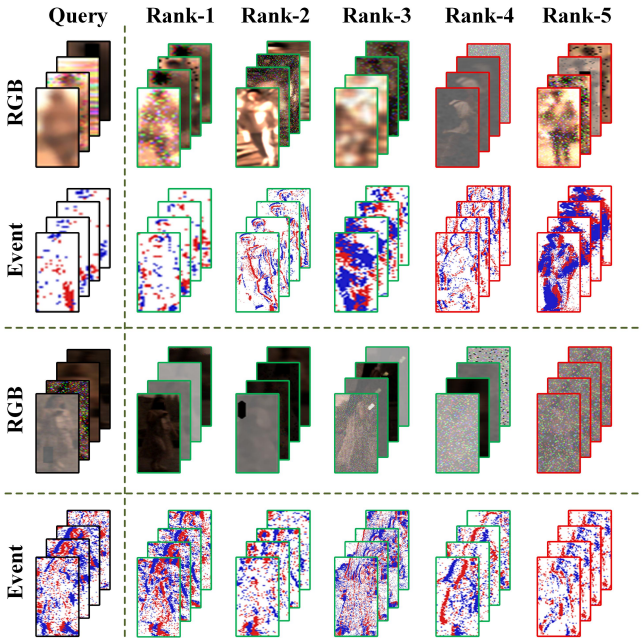

Figure 4: Visualization of the rank list.

tary 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 TriPro-ReID on the MARS* dataset using both RGB and Event modalities. As a large-scale and challenging benchmark, MARS* provides a rigorous testbed for assessing model robustness. As shown in Table 3, our method achieves 88.4% mAP and 91.1% Rank-1 accuracy, surpassing all existing RGB-Event based methods. This demonstrates the effectiveness of our framework in learning discriminative spatiotemporal features. While not outperforming some unimodal methods, this is likely due to the synthetic nature of the Event modality in MARS*, which may introduce noise and hinder full modality complementarity.

Ablation Studies

In this section, we first present the results evaluating the effectiveness of key components. We further provide detailed experimental analyses of different settings within each component. Additional results and implementation details are available in the supplementary materials.

• **Effect of Key Components.** As shown in Table 4, adding PNAP or CMP individually to the base model leads to consistent performance improvements on both EvReID and MARS* datasets. Introducing PNAP increases the mAP from 49.2% to 62.3% on EvReID, showing its effectiveness in modality-degraded scenarios. CMP also brings gains by enhancing cross-modal interaction. When both modules are used together, the model achieves the best results, demonstrating the complementary advantages of PNAP and CMP.

• **Effect of Different Depth and Length of CMP.** As shown in Table 5, shorter prompt lengths outperform longer ones, with the best mAP of 69.8% achieved at length 10. For depth, the best performance is observed at depth 12, while reducing it to 6 or 3 leads to noticeable drops in accuracy.

Visualization of Rank List

Fig. 4 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.

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 (Wang et al. 2023) will be explored to enhance the overall accuracy.

Acknowledgments

This work is supported in part by Grant No. 2023-JCJQ-LA-001-088 and in part by Grant No. 2025ZD1601300. It is also supported by the National Natural Science Foundation of China under Grant No. 62102205 and U24A20342, the Anhui Provincial Natural Science Foundation–Outstanding Youth Project (2408085Y032), the Natural Science Foundation of Anhui Province (2408085J037), and the Key Technologies R&D Program of Anhui Province (202423k09020039). The authors also acknowledge the High-Performance Computing Platform of Anhui University for providing computational resources.

References

Ahmad, S.; Morerio, P.; and Del Bue, A. 2023. Person Re-Identification without Identification via Event anonymization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 11132–11141.

Bai, S.; Ma, B.; Chang, H.; Huang, R.; and Chen, X. 2022. Salient-to-Broad Transition for Video Person Re-identification. In *CVPR*.

Bialkowski, A.; Denman, S.; Sridharan, S.; Fookes, C.; and Lucey, P. 2012. A database for person re-identification in multi-camera surveillance networks. In *2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*, 1–8. IEEE.

Cao, C.; Fu, X.; Liu, H.; Huang, Y.; Wang, K.; Luo, J.; and Zha, Z.-J. 2023. Event-guided person re-identification via sparse-dense complementary learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17990–17999.

Chen, D.; Doering, A.; Zhang, S.; Yang, J.; Gall, J.; and Schiele, B. 2022. Keypoint message passing for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 239–247.

Cheng, Y.; Knoll, A.; and Cao, H. 2025. UR-Net: uncertainty-aware refinement network for event-based stereo depth estimation. *Visual Intelligence*, 3(1): 18.

Deng, Y.; Chen, Z.; Li, C.; and Tang, J. 2025a. Uncertainty-aware coarse-to-fine alignment for text-image person retrieval. *Visual Intelligence*, 3(1): 6.

Deng, Y.; Li, C.; Chen, Z.; Xu, Z.; and Tang, J. 2025b. Decoupled Cross-Modal Alignment Network for Text-RGBT Person Retrieval and A High-Quality Benchmark. *Information Fusion*, 103948.

Deng, Y.; Li, C.; Wang, F.; and Tang, J. 2025c. Learning Hierarchical Cross-modal Association with Intra-modal Context for Text-Image Person Retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2723–2731.

Eom, C.; Lee, G.; Lee, J.; and Ham, B. 2021. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12036–12045.

Gehrig, M.; and Scaramuzza, D. 2023. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13884–13893.

Gu, X.; Chang, H.; Ma, B.; Zhang, H.; and Chen, X. 2020. Appearance-Preserving 3D Convolution for Video-based Person Re-identification. In *ECCV*.

Hirzer, M.; Beleznai, C.; Roth, P. M.; and Bischof, H. 2011. Person re-identification by descriptive and discriminative classification. In *Image Analysis: 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, May 2011. Proceedings 17*, 91–102. Springer.

Hou, R.; Chang, H.; Ma, B.; Huang, R.; and Shan, S. 2021. BiCnet-TKS: Learning Efficient Spatial-Temporal Representation for Video Person Re-Identification. In *CVPR*.

Hou, R.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2020. Temporal Complementary Learning for Video Person Re-Identification. In *ECCV*.

Huang, J.; Wang, S.; Wang, S.; Wu, Z.; Wang, X.; and Jiang, B. 2024. Mamba-fetrack: Frame-event tracking via state space model. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 3–18. Springer.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European conference on computer vision*, 709–727. Springer.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19113–19122.

Li, J.; Wang, J.; Tian, Q.; Gao, W.; and Zhang, S. 2019. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3958–3967.

Li, R.; Yuan, X.; Liu, W.; and Xu, X. 2025. Event-based Video Person Re-identification via Cross-Modality and Temporal Collaboration. *arXiv preprint arXiv:2501.07296*.

Li, S.; Sun, L.; and Li, Q. 2023. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 1405–1413.

Liu, J.; Zha, Z.-J.; Wu, W.; Zheng, K.; and Sun, Q. 2021a. 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*, 4370–4379.

Liu, X.; Yu, C.; Zhang, P.; and Lu, H. 2023. Deeply coupled convolution–transformer with spatial–temporal complementary learning for video-based person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*.

Liu, X.; Zhang, P.; Yu, C.; Lu, H.; and Yang, X. 2021b. 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*, 13334–13343.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Nambiar, A.; Taiana, M.; Figueira, D.; Nascimento, J. C.; and Bernardino, A. 2014. A multi-camera video dataset for research on high-definition surveillance. *International Journal of Machine Intelligence and Sensory Signal Processing*, 1(3): 267–286.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Shu, X.; Wang, X.; Zang, X.; Zhang, S.; Chen, Y.; Li, G.; and Tian, Q. 2021. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4390–4403.

Wang, H.; Jiao, L.; Yang, S.; Li, L.; and Wang, Z. 2020. Simple and effective: Spatial rescaling for person reidentification. *IEEE Transactions on neural networks and learning systems*, 33(1): 145–156.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person re-identification by video ranking. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, 688–703. Springer.

Wang, X.; Chen, G.; Qian, G.; Gao, P.; Wei, X.-Y.; Wang, Y.; Tian, Y.; and Gao, W. 2023. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4): 447–482.

Wang, X.; Jin, Y.; Wu, W.; Zhang, W.; Zhu, L.; Jiang, B.; and Tian, Y. 2024a. Object Detection using Event Camera: A MoE Heat Conduction based Detector and A New Benchmark Dataset. arXiv:2412.06647.

Wang, X.; Zheng, S.; Yang, R.; Zheng, A.; Chen, Z.; Tang, J.; and Luo, B. 2022. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121: 108220.

Wang, X.; Zhu, Q.; Jin, J.; Zhu, J.; Wang, F.; Jiang, B.; Wang, Y.; and Tian, Y. 2024b. Spatio-temporal side tuning pre-trained foundation models for video-based pedestrian attribute recognition. *arXiv preprint arXiv:2404.17929*.

Wang, Y.; Liu, Y.; Zheng, A.; and Zhang, P. 2025a. Decoupled feature-based mixture of experts for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8141–8149.

Wang, Y.; Lv, Y.; Zhang, P.; and Lu, H. 2025b. 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)*.

Wang, Y.; Zhang, P.; Gao, S.; Geng, X.; Lu, H.; and Wang, D. 2021. Pyramid Spatial-Temporal Aggregation for Video-based Person Re-Identification. In *ICCV*.

Wu, J.; Huang, Y.; Gao, M.; Niu, Y.; Chen, Y.; and Wu, Q. 2025a. Enhanced Visual-Semantic Interaction with Tailored Prompts for Pedestrian Attribute Recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9570–9579.

Wu, J.; Huang, Y.; Gao, M.; Niu, Y.; Chen, Y.; Wu, Q.; and Zhao, J. 2025b. Learning comprehensive representation via selective activation and dual-level orthogonality for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wu, J.; Huang, Y.; Gao, M.; Niu, Y.; Yang, M.; Gao, Z.; and Zhao, J. 2024. Selective and orthogonal feature activation for pedestrian attribute recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 6039–6047.

Wu, Y.; Lin, Y.; Dong, X.; Yan, Y.; Ouyang, W.; and Yang, Y. 2018. 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*, 5177–5186.

Xu, Y.; Wu, M.; Guo, Z.; Cao, M.; Ye, M.; and Laaksonen, J. 2025. Efficient text-to-video retrieval via multi-modal multi-tagger derived pre-screening. *Visual Intelligence*, 3(1): 1–13.

Yan, Y.; Qin, J.; Chen, J.; Liu, L.; Zhu, F.; Tai, Y.; and Shao, L. 2020. 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*, 2896–2905. IEEE.

Yu, C.; Liu, X.; Wang, Y.; Zhang, P.; and Lu, H. 2024. TF-CLIP: Learning text-free CLIP for video-based person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 6764–6772.

Yu, C.; Liu, X.; Zhu, J.; Wang, Y.; Zhang, P.; and Lu, H. 2025. Climb-reid: A hybrid clip-mamba framework for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9589–9597.

Zhang, Q.; Wang, L.; Patel, V. M.; Xie, X.; and Lai, J. 2024. 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*, 22000–22009.

Zheng, A.; Liu, J.; Wang, Z.; Huang, L.; Li, C.; and Yin, B. 2023. Visible-infrared person re-identification via specific and shared representations learning. *Visual Intelligence*, 1(1): 29.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. 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*, 868–884. Springer.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, K.; Yang, Y.; Cavallaro, A.; and Xiang, T. 2019. Omni-Scale Feature Learning for Person Re-Identification. arXiv:1905.00953.