Dynamic-Static Collaboration for Unsupervised Domain Adaptive Video-Based Visible-Infrared Person Re-Identification

Jiaxu Leng[1,2], Zhengjie Wang[1,2], Shuang Li[1,2]*, Xinbo Gao[1,2]*

[1]School of Computer Science and Technology, Chongqing University of Posts and Telecommunications
[2]Chongqing Institute for Brain and Intelligence, Guangyang Bay Laboratory, Chongqing, China
lengjx@cqupt.edu.cn, s240201156@stu.cqupt.edu.cn, shuangli936@gmail.com, gaoxb@cqupt.edu.cn

Abstract

Video-based visible-infrared person re-identification (VVI-ReID) aims to match pedestrian sequences across modalities for all-day surveillance. While supervised methods have shown progress, their dependence on large-scale cross-modal annotations limits scalability. We investigate the task of unsupervised domain adaptation for VVI-ReID (UDA-VVI-ReID), where a model trained on a labeled source domain is adapted to an unlabeled target domain. Directly extending existing image-based unsupervised VI-ReID methods to video scenarios by simply averaging frame-level features is suboptimal, as this naive strategy neglects the rich temporal dynamics in video data and leads to unreliable pseudo-labels due to occlusion-induced noise. To overcome these limitations, we propose a Dynamic-Static Collaboration (DSC) framework that explicitly leverages the complementary strengths of motion and appearance cues. The Dynamic-Static Label Unification (DSLU) module refines pseudo-labels by validating the consistency between static and dynamic predictions. Based on these labels, the Dynamic-Static Joint Learning (DSJL) module performs neighbor-aware contrastive learning in both feature spaces, promoting robust representation learning under cross-modal and temporal variations. Experiments on HITSZ-VCM and BUPTCampus show that DSC sets a strong baseline for this new task, enabling robust cross-modal video ReID without target labels.

Code — https://github.com/YwAcle/DSC

Introduction

Video-based Visible-Infrared Person Re-Identification (VVI-ReID) aims to retrieve pedestrian video sequences across different modalities under varying viewpoints. This task plays a vital role in intelligent surveillance systems, particularly under challenging conditions such as day-night transitions and low-light environments (Lin et al. 2022; Du et al. 2023; Zhang and Wang 2023; Wei et al. 2021). Recent supervised approaches have achieved notable progress by leveraging large-scale annotated video datasets (Lin et al. 2022; Du et al. 2023). However, annotating cross-modal identity associations at the video level is labor-intensive and time-consuming, severely limiting scalability in real-world deployments.

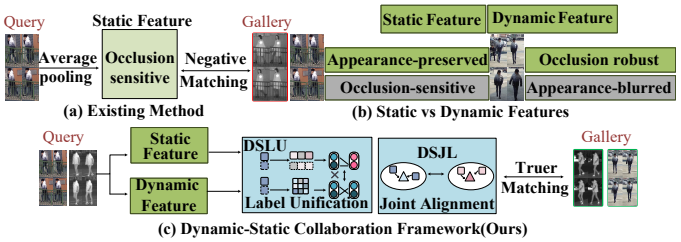

Figure 1: Overview of motivation and the proposed framework. (a) Existing methods rely on average pooling for frame aggregation, which is vulnerable to occlusion and modality shifts. (b) Dynamic features offer occlusion robustness but lack appearance specificity, while static features preserve texture but are sensitive to noise—revealing a trade-off. (c) Our method addresses this by enforcing consistency between dynamic and static cues via Dynamic-Static Label Unification (DSLU) and Dynamic-Static Joint Learning (DSJL), which collaboratively refine pseudo-labels and enhance cross-modal alignment.

To alleviate the reliance on manual annotations, we draw inspiration from Unsupervised Domain Adaptation (UDA) methods in person re-identification (ReID) (Ge, Chen, and Li 2020; Ge et al. 2020). These methods typically perform supervised pretraining on a labeled source domain and adopt clustering-based pseudo-label generation for unsupervised learning in the target domain. In this work, we extend this paradigm to the video-based visible-infrared setting and formulate the task of Unsupervised Domain Adaptation for Visible-Infrared Video Re-Identification (UDA-VVI-ReID), where a source-pretrained model is adapted to unlabeled video sequences in the target domain. This approach aims to enhance generalization under modality discrepancies and environmental variations, without requiring target domain labels.

A straightforward approach to adapting image-based unsupervised VI-ReID (USL-VI-ReID) methods for the UDA-VVI-ReID task is to apply average pooling to aggregate frame-level features into a sequence-level representation, thereby enabling direct reuse of USL-VI-ReID techniques

*Corresponding authors.

in the video-based cross-modal setting (Cheng et al. 2023a; Wu and Ye 2023; Shi et al. 2023; Ye, Shen, and Shao 2020). However, as illustrated in Fig. 1(a), this naive extension suffers under persistent occlusions and modality shifts, leading to severely corrupted pseudo-labels. In contrast, temporal modeling (e.g., Fig. 1(b)) captures motion continuity and improves robustness to occlusion, but often overlooks fine-grained appearance details such as clothing texture—features that are crucial for cross-modal identity discrimination. These limitations reveal a fundamental tradeoff between temporal robustness and appearance discriminability. Static features encode rich identity cues that are stable across modalities but sensitive to occlusion and motion blur. Dynamic features, derived from motion trajectories or pose transitions, are more resilient in challenging environments but often lack semantic specificity for fine-grained identity matching. Moreover, this discrepancy is further amplified in visible-infrared settings, where appearance and motion are encoded differently due to modality-induced distortions. To tackle this, we argue that dynamic and static features should not be modeled independently nor simply fused, but instead used to validate each other. Specifically, we observe that when both cues independently predict the same label, the result is more likely to be correct. This motivates our consistency-driven pseudo-label refinement strategy, where only predictions supported by both dynamic and static views are retained. By enforcing cross-view agreement, we suppress unreliable labels and enhance the discriminability of learned representations.

To realize dynamic-static consistency, we propose a unified framework that jointly performs pseudo-label refinement and representation learning in a mutually reinforcing manner as shown in Fig. 1(c). It comprises two key modules tailored to address label noise and modality discrepancy under unsupervised settings. The Dynamic-Static Label Unification (DSLU) module enhances pseudo-label reliability by enforcing consistency between motion and appearance predictions across modalities. Instead of relying on single-view clustering, DSLU independently extracts dynamic and static features from video sequences, retaining labels only when both views agree—effectively filtering out noise caused by occlusion, blur, or cross-modal distortion. Building on this supervision, the Dynamic-Static Joint Learning (DSJL) module improves representation learning via neighbor-aware contrastive learning. By jointly considering dynamic and static similarities, DSJL adaptively selects reliable intra- and inter-modality neighbors, promoting robust alignment and mitigating domain shifts. Together, DSLU and DSJL form a closed-loop learning paradigm: the former enhances supervision quality, while the latter progressively improves feature discriminability, thereby reinforcing the effectiveness of pseudo-label refinement.

Our contributions are as follows:(1) We formulate the **UDA-VVI-ReID** task, enabling cross-modal video person re-identification without target domain labels.(2) We propose the **DSLU** module to generate reliable pseudo-labels by enforcing dynamic-static consistency across modalities.(3) We design the **DSJL** module to enhance feature learning via contrastive and neighbor-based objectives guided by both motion and appearance cues.(4) Extensive experiments on two challenging benchmarks, **HITSZ-VCM** and **BUPT-Campus**, demonstrate that our method achieves state-of-the-art performance among unsupervised approaches and even outperforms several supervised baselines.

The rest of this paper is organized as follows. Section II introduces related work;Section III elaborates the proposed method; Section IV analyzes the comparative experimental results; and Section V concludes this paper.

Related Work

Supervised Video-based Visible-Infrared Person Re-Identification

Video-based Visible-Infrared Person Re-Identification (VVI-ReID) aims to match pedestrian sequences across visible and infrared modalities by exploiting spatiotemporal cues for robust identity recognition in challenging scenarios like low illumination and day-night transitions (Ristani et al. 2016; Wu et al. 2017).

Early efforts mainly focused on image-level VI-ReID via feature alignment. For example, Ye et al. (Ye et al. 2021a) adopted graph attention to model multi-level relations, while Lin et al. (Lin et al. 2022) introduced modality-invariant temporal memory to extract dynamic features. However, lacking temporal modeling, such methods struggle in video-level tasks.

Recent approaches incorporate motion modeling for cross-modal video matching. Du et al. (Du et al. 2023) enhanced robustness via auxiliary samples, Cui et al. (Cui, Zhou, and Peng 2024) proposed dual-modality alignment, and Li et al. (Li et al. 2025) used language-driven prompts and CLIP-based representations for lightweight spatiotemporal alignment. Despite promising results, these methods often overlook fine-grained static cues (e.g., clothing texture) critical for discriminative representation.

Some recent works explore combining static and dynamic features (Chen et al. 2022; Ye et al. 2023), yet effectively integrating these heterogeneous cues for reliable pseudo-labeling and robust retrieval remains challenging.

Unsupervised Visible-Infrared Image Person Re-Identification

Unsupervised VI-ReID (USL-VI-ReID) aims to enable cross-modal matching without labels, improving scalability over supervised methods. Early approaches adopt contrastive learning and clustering. ADCA (Yang et al. 2022a) introduces dual-contrastive aggregation, HDSA (Zhang et al. 2022a) applies hierarchical spatial alignment, and Cluster Contrast (Dai et al. 2022) enhances modality invariance via cluster-based contrast.

Subsequent methods improve pseudo-label quality and semantic consistency. SALCR (Cheng et al. 2023b) uses attention-guided refinement, CHCR (Pang et al. 2023) leverages hierarchical clustering, and PGM+AL (Wu and Ye 2023) combines graph matching and alternate training. SDCL (Yang, Chen, and Ye 2024) introduces shallow-deep collaboration with Transformers, achieving competitive results.

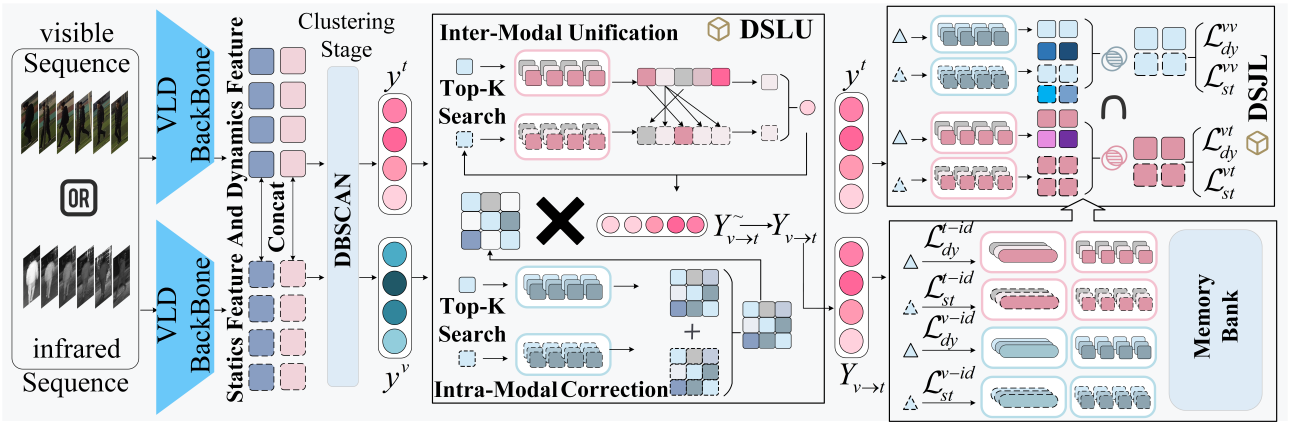

Figure 2: The proposed DSC framework consists of two modules: Dynamic-Static label unification (DSLU) and Dynamic-Static joint learning (DSJL). By utilizing complementary information at both the Dynamic and Static levels, the pseudo labels of the two modalities are unified at the clustering level to jointly train the information of the two modalities with the highest confidence, promoting cross modal fusion. DSJL uses two complementary information to find reliable neighboring samples at the sample level for convergence, enhancing the robustness of unsupervised learning

While effective, most USL-VI-ReID methods prioritize static appearance alignment and ignore motion information crucial in video scenarios. Thus, integrating temporal dynamics with static cues under unsupervised settings remains a key challenge for advancing video-level VI-ReID.

Method

Our method first performs supervised pretraining on a labeled source domain, and then refines pseudo-labels via dynamic-static consistency in the **DSLU** module, followed by contrastive and neighbor-aware learning in the **DSJL** module.

Unsupervised Domain Adaptive Video-Based Dual-Contrastive Learning Framework

Source-Domain Pretraining with Spatio-Temporal Prompting To extract modality-invariant yet discriminative representations, we perform supervised pretraining on the source domain with the Spatial-Temporal Prompting (STP) module (Li et al. 2025), which explicitly decouples appearance and motion information. Given a visible-infrared video dataset $\mathcal{T} = \{\mathcal{T}^v, \mathcal{T}^r\}$, where $\mathcal{T}^v$ and $\mathcal{T}^r$ denote visible and infrared sequences respectively, each input video $V \in \mathbb{R}^{T\times H\times W\times C}$ is tokenized into patches with positional encodings and fed into a ViT backbone. A learnable spatio-temporal prompt tensor $z_p \in \mathbb{R}^{T\times T\times D}$ is appended to the token sequence and processed across layers with alternating temporal-axis transposition. This enables the model to capture localized spatial patterns and global temporal dependencies. After several layers of interaction, we extract two complementary sequence-level representations: (1) a static feature f_{st} via average pooling of CLS tokens, and (2) a dynamic feature f_{dy} computed by querying the prompt tokens through multi-head attention (MHA). To encourage identity separability, both f_{st} and f_{dy} are supervised using cross-entropy and triplet losses (Ye et al. 2021c). This pretraining provides initialization for subsequent adaptation.

Dual-Contrastive Learning in the Target Domain To adapt the pretrained model to the unlabeled target domain, we propose a dual-branch contrastive learning framework inspired by ADCA (Yang et al. 2022a), which jointly models static and dynamic representations. For each modality, taking the visible modality as an example, the concatenated features $\left[f^v_{st}, f^v_{dy}\right] \in \mathbb{R}^{N_v\times 2d}$ are clustered using DBSCAN, yielding cluster assignments $\{\mathcal{I}^v_1, \ldots, \mathcal{I}^v_{C_v}\}$. The pseudo-labels for the visible and infrared modalities are denoted as y^v_i and y^t_i, respectively, indicating the assigned cluster index for sample i in each modality.For each cluster c, the static and dynamic centers are computed as:

$$\mu^v_{dy,c} = \frac{1}{|\mathcal{I}^v_c|}\sum_{i\in\mathcal{I}^v_c} f^v_{dy,i}, \quad \mu^v_{st,c} = \frac{1}{|\mathcal{I}^v_c|}\sum_{i\in\mathcal{I}^v_c} f^v_{st,i}, \tag{1}$$

where $|\cdot|$ denotes the cardinality operator, i.e., the number of elements in a set. These cluster centers are stored in modality-specific memory banks, denoted as $\mathcal{M}^v_{dy}$ and $\mathcal{M}^v_{st}$, which serve as dynamic and static anchors for contrastive learning. To maintain stable and representative prototypes during training, we adopt a momentum-based update strategy. For a feature q from the current mini-batch assigned to cluster k, the corresponding memory entry is updated as:

$$\mathcal{M}^v_{(\cdot)}[k] \leftarrow \beta\cdot\mathcal{M}^v_{(\cdot)}[k] + (1-\beta)\cdot q, \tag{2}$$

where $\beta \in [0, 1)$ is the momentum coefficient controlling the update rate, and $(\cdot) \in \{dy, st\}$ denotes the dynamic or static branch. Given a sample feature (e.g., f^v_{dy}), it is pulled towards its positive center and pushed away from negatives:

$$\mathcal{L}^{id,v}_{dy} = -\log\frac{\exp\left(f^v_{dy}\cdot\mathcal{M}^v_{dy,+}/\tau\right)}{\sum_{k=1}^{K}\exp\left(f^v_{dy}\cdot\mathcal{M}^v_{dy,k}/\tau\right)}, \tag{3}$$

$$\mathcal{L}_{st}^{id,v} = -\log \frac{\exp\left(f_{st}^{v} \cdot \mathcal{M}_{st,+}^{v}/\tau\right)}{\sum_{k=1}^{K} \exp\left(f_{st}^{v} \cdot \mathcal{M}_{st,k}^{v}/\tau\right)}, \quad (4)$$

where τ is a temperature parameter and K is the number of negatives. Analogous losses $\mathcal{L}_{dy}^{id,t}$ and $\mathcal{L}_{st}^{id,t}$ are computed for the infrared modality.

Dynamic-Static Label Unification

While dual-contrastive frameworks aid target-domain adaptation, they often neglect the complementary roles of static and dynamic features, leading to suboptimal pseudo-label quality. To address this, we propose the DSLU module, which refines pseudo-labels by jointly exploiting static features f_{st} and dynamic features f_{dy}. DSLU is motivated by two observations: (1) Intra-modal clustering suffers from noise and fragmentation under occlusion or viewpoint shifts, which can be alleviated by enforcing consistency between f_{st} and f_{dy} within each modality; (2) Despite the significant domain gap between visible and infrared modalities, agreement between f_{st} and f_{dy} across modalities offers robust identity cues. Accordingly, DSLU refines pseudo-labels in two stages: intra-modal correction, which rectifies clustering errors via within-modality consistency, and inter-modal unification, which aligns labels across modalities based on cross-feature agreement.

Inter-Modal Label Unification To enhance pseudo-label consistency across modalities, we propose an inter-modal label unification strategy that jointly leverages static and dynamic features to mitigate noise caused by modality discrepancies. Specifically, given a visible sample x_i^v and infrared samples $x_j^t \in \{x_1^t, \ldots, x_{N_t}^t\}$, from which we extract static and dynamic features $f_{st,i}^v, f_{dy,i}^v \in \mathbb{R}^d$ and $f_{st,j}^t, f_{dy,j}^t \in \mathbb{R}^d$. We then compute the cosine similarity between the visible query and all infrared samples in both feature spaces:

$$sim_{dy}^{v\to t} = \frac{f_{dy}^{v} \cdot (f_{dy}^{t})^{\top}}{\|f_{dy}^{v}\|\|f_{dy}^{t}\|}, \quad sim_{st}^{v\to t} = \frac{f_{st}^{v} \cdot (f_{st}^{t})^{\top}}{\|f_{st}^{v}\|\|f_{st}^{t}\|}, \quad (5)$$

where $sim_{dy}^{v\to t}, sim_{st}^{v\to t} \in \mathbb{R}^{N_t}$ are the similarity vectors between the visible query and all infrared samples in the dynamic and static feature spaces, respectively. Then we perform top-K retrieval in both spaces to select the most similar infrared samples:

$$I_{dy} = TopK(sim_{dy}^{v\to t}), \quad I_{st} = TopK(sim_{st}^{v\to t}), \quad (6)$$

where $I_{dy}, I_{st} \in \mathbb{N}^K$ denote the indices of the top-K most similar infrared samples ranked by dynamic and static similarity, respectively. Based on the dynamic similarity ranking, we extract the candidate pseudo-labels from the infrared label set $y^t \in \mathbb{N}^{N_I}$:

$$Y_{cand} = y^t[I_{dy}], \quad (7)$$

where $Y_{cand} \in \mathbb{N}^K$ denotes the label candidates associated with the dynamic top-K neighbors. To enforce dynamic-static consistency, we measure the agreement between dynamic and static neighbors:

$$I_k = \sum_{i=1}^{K} \mathbb{I}(Y_{cand} = y^t[I_{st}][:, i]), \quad (8)$$

where $\mathbb{I}(\cdot)$ is the indicator function, and $I_k \in \mathbb{N}^K$ counts how many times each candidate label appears among the static top-K neighbors. Finally, we adopt a majority voting strategy to assign the unified pseudo-label:

$$\hat{Y} = \arg\max_k I_k, \quad (9)$$

where $\hat{Y}$ is the final pseudo-label selected based on the highest agreement between dynamic and static rankings. This inter-modal label unification strategy filters out inconsistent cross-modal predictions by enforcing agreement across motion and appearance spaces, thus yielding more robust pseudo-labels for subsequent learning.

Intra-Modal Label Correction Pseudo-labels within each modality are often corrupted by occlusions and viewpoint variations. To mitigate this, we exploit the complementary nature of static (appearance) and dynamic (motion) features to enhance label consistency. Starting with static features $f_{st}^V \in \mathbb{R}^{N\times d}$ and dynamic features $f_{dy}^V \in \mathbb{R}^{N\times d}$, we compute their respective intra-modal cosine similarities as:

$$sim_{\text{st}}^{v} = \frac{f_{\text{st}}^{v} \cdot (f_{\text{st}}^{v})^{\top}}{\|f_{\text{st}}^{v}\|^2}, \quad sim_{\text{dy}}^{v} = \frac{f_{\text{dy}}^{v} \cdot (f_{\text{dy}}^{v})^{\top}}{\|f_{\text{dy}}^{v}\|^2}, \quad (10)$$

Here, $(\cdot)^{\top}$ denotes the matrix transpose, and $\|\cdot\|$ represents the row-wise ℓ_2-norm used for normalization. To jointly encode appearance and motion cues, the static and dynamic similarity matrices are fused into a unified similarity matrix:

$$sim^{v} = \lambda_{\text{st}} \cdot sim_{\text{st}}^{v} + \lambda_{\text{dy}} \cdot sim_{\text{dy}}^{v}, \quad (11)$$

where $\lambda_{\text{st}}, \lambda_{\text{dy}} \in [0, 1]$ control the contribution of each modality. To incorporate local structural information for label refinement, the top-K most similar neighbors are selected for each sample based on the fused similarity matrix:

$$I = TopK(\text{sim}^{v}), \quad (12)$$

where $TopK$ selects the indices of the top-K most similar samples for each instance based on the fused similarity matrix sim^v. To facilitate label aggregation based on local neighborhood structure, a binary mask $M \in \{0, 1\}^{N\times N}$ is constructed to indicate neighbor relationships:

$$M = scatter(I, 1), \quad (13)$$

where $scatter$ assigns 1s to the positions of selected neighbors. The final corrected label distribution $\mathbf{Y}_{v->t} \in \mathbb{R}^{N\times C}$ is computed by aggregating the neighbor labels $\hat{Y} \in \mathbb{R}^{N\times C}$:

$$\mathbf{Y}_{v\to t} = M\hat{Y}, \quad (14)$$

where $\hat{Y} \in \mathbb{R}^{N_V\times C}$ is the unified pseudo-label matrix obtained from **Inter-Modal Label Unification**, and $\mathbf{Y}_{v->t}$ denotes the refined intra-modal label distribution.This process enforces dynamic-static consistency at the neighborhood level and suppresses intra-modal noise through local consensus.

Dynamic-Static Joint Learning (DSJL)

While DSLU enhances pseudo-label reliability through dynamic-static consistency, robust representation learning remains hindered by two critical limitations: (1) the over-reliance on top-1 pseudo-label matches, which are sensitive to modality shifts and occlusions; and (2) the exclusion of outlier samples (with pseudo-label -1) that are not utilized due to clustering failures. To mitigate these issues, we propose the **Dynamic-Static Joint Learning (DSJL)** module, which promotes robust learning by enforcing cross-modal neighbor consensus under both motion and appearance cues. Rather than relying on a single top-1 match, DSJL identifies multi-view consistent neighbors to form adaptive positive sets, enabling stable and noise-tolerant supervisio Specifically, taking a query sample q_i as an example, we compute its cosine similarity with all training samples in both dynamic and static spaces:

$$s_{dy}(q_i, u_j) = \frac{f_{dy,i} \cdot f_{dy,j}}{\|f_{dy,i}\|\|f_{dy,j}\|}, \tag{15}$$

$$s_{st}(q_i, u_j) = \frac{f_{st,i} \cdot f_{st,j}}{\|f_{st,i}\|\|f_{st,j}\|}, \tag{16}$$

where $f_{dy,i}, f_{st,i} \in \mathbb{R}^d$ denote the dynamic and static features of query q_i, and u_j denotes any training sample.To adaptively control neighbor selection, we define thresholds based on the maximum similarity for each query:

$$\theta_i^{dy} = \alpha \cdot \max_j s_{dy}(q_i, u_j), \tag{17}$$

$$\theta_i^{st} = \alpha \cdot \max_j s_{st}(q_i, u_j), \tag{18}$$

where $\alpha \in (0, 1)$ is a scaling factor that adjusts the neighbor sensitivity, and $\theta_i^{dy}, \theta_i^{st} \in \mathbb{R}$ are the dynamic and static similarity thresholds for sample q_i.Based on these thresholds, soft neighbor sets are selected as:

$$N_{dy}(q_i) = \{u_j \mid s_{dy}(q_i, u_j) > \theta_i^{dy}\}, \tag{19}$$

$$N_{st}(q_i) = \{u_j \mid s_{st}(q_i, u_j) > \theta_i^{st}\}, \tag{20}$$

where $N_{dy}(q_i), N_{st}(q_i)$ are the neighbor sets selected based on dynamic and static similarity thresholds, respectively.To ensure both motion and appearance consistency, the final neighbor set is defined as their intersection:

$$N(q_i) = N_{dy}(q_i) \cap N_{st}(q_i), \tag{21}$$

where $N(q_i)$ contains neighbors that are similar to the query in both feature spaces. Given the selected neighbors, DSJL applies neighbor-aware contrastive learning to enforce alignment across and within modalities. For example, the visible-to-infrared loss is defined as:

$$\mathcal{L}^{vt} = \mathcal{L}_{dy}^{vt} + \mathcal{L}_{st}^{vt}, \tag{22}$$

$$\mathcal{L}_{dy}^{vt} = -\log \frac{\exp\left(f_{dy}^{v} \cdot f_{dy,+}^{t}/\tau\right)}{\sum_{k=0}^{K} \exp\left(f_{dy}^{v} \cdot f_{dy,k}^{t}/\tau\right)}, \tag{23}$$

$$\mathcal{L}_{st}^{vt} = -\log \frac{\exp\left(f_{st}^{v} \cdot f_{st,+}^{t}/\tau\right)}{\sum_{k=0}^{K} \exp\left(f_{st}^{v} \cdot f_{st,k}^{t}/\tau\right)}, \tag{24}$$

where $f_{dy,+}^{t}$, $f_{st,+}^{t}$ are positive infrared samples, and $f_{dy,k}^{t}$, $f_{st,k}^{t}$ are sampled negatives. Similar losses $\mathcal{L}^{vv}$, $\mathcal{L}^{tt}$, and $\mathcal{L}^{tv}$ are computed for intra- and reverse cross-modal relations.

Optimization

In training process, we apply identity contrastive losses $\mathcal{L}_{dy}^{id}$ and $\mathcal{L}_{st}^{id}$ to supervise dynamic and static branches, respectively, ensuring discriminative representations. To facilitate robust cross-modal alignment and enhance feature consistency guided by dynamic-static relations, we further adopt neighbor losses $\mathcal{L}^{vt}$, $\mathcal{L}^{vv}$, $\mathcal{L}^{tt}$, and $\mathcal{L}^{tv}$ across and within modalities.The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{dy}^{id} + \mathcal{L}_{st}^{id} + \lambda_1(\mathcal{L}^{vv} + \mathcal{L}^{tt}) + \lambda_2(\mathcal{L}^{vt} + \mathcal{L}^{tv}), \tag{25}$$

where λ_1 and λ_2 are weighting factors that balance intra-modal and cross-modal neighbor losses, respectively.

Experiments

Datasets and Evaluation Metrics. We evaluate our method on two large-scale video-based visible-infrared ReID datasets: HITSZ-VCM (Lin et al. 2022) and BUPTCampus (Du et al. 2023). HITSZ-VCM contains 927 identities with 251,452 RGB and 211,807 IR images across 24-frame tracklets, split into 500 identities for training and 427 for testing. BUPTCampus comprises 3,080 identities and 1.87M frames across 16,826 tracklets, divided into 1,074/930/1,076 identities for primary/auxiliary/testing sets. We follow standard evaluation using CMC and mAP, and report performance in both RGB↔IR retrieval directions.

Implementation Details

Our framework is implemented in PyTorch and trained on a single A6000 GPU with seed 1 for reproducibility. Input images are resized to 288×144 and augmented via random flip, padding, and cropping. We adopt CLIP-ViT-16 as the image encoder and VLD as the backbone. Optimization uses Adam with a 0.00025 base learning rate, cosine decay, and 50 training epochs. Batch sizes are 16 (IR) and 8 (RGB); clustering uses batch size 128 with 4 positives per identity per modality. Training adopts random identity and consistent modality sampling, supervised by contrastive and neighborhood losses for cross- and intra-modal alignment.

Comparative Experimental Analysis

We evaluate the proposed DSC framework on two public video-based visible-infrared person re-identification (VVI-ReID) datasets: HITSZ-VCM and BUPTCampus. Following prior works, we report Rank-1 and mean Average Precision (mAP). We compare DSC with both unsupervised methods (ADCA (Yang et al. 2022a), PGM (Wu and Ye 2023), GUR (Yang, Chen, and Ye 2023), NG (Cheng et al. 2023b)) and supervised methods (DDAG (Ye et al. 2020), LbA (Park et al. 2021), CAJ (Ye et al. 2021b), AGW (Ye et al. 2021c), MMN (Zhang et al. 2021) DART (Yang et al. 2022b), DEEN (Zhang and Wang 2023), CLIP-ReID (Li, Sun, and Li 2023), TF-CLIP (Yu et al. 2024),MITML(Lin et al. 2022), IBAN (Li et al. 2023a), SADSTRM (Li et al. 2023b), SAADG (Zhou et al. 2023), CST (Feng et al. 2024),AuxNet (Du et al. 2023)).

As shown in Table **??** and Table **??**, DSC significantly outperforms existing unsupervised approaches in both RGB→IR and IR→RGB directions. Moreover, it

	Method	Reference	RGB→IR		IR→RGB	
			Rank-1	mAP	Rank-1	mAP
Supervised	DDAG	ECCV'20	40.4	40.4	46.3	43.1
	LbA	ICCV'21	32.1	32.9	39.1	37.1
	CAJ	ICCV'21	40.5	41.5	45.0	43.6
	AGW	TPAMI'21	36.4	37.4	43.7	41.1
	MMN	CVPR'21	40.9	41.7	43.7	42.8
	DART	CVPR'22	52.4	49.1	53.3	50.5
	DEEN	CVPR'23	53.7	50.4	49.8	48.6
	CLIP-ReID	AAAI'23	49.0	50.4	51.0	49.8
	TF-CLIP	AAAI'24	49.4	51.9	52.5	51.8
	MITML	CVPR'22	49.1	47.5	50.2	46.3
Unsupervised	CD	-	25.2	26.8	19.2	19.9
	ADCA	ACM MM'22	29.6	29.1	27.6	27.8
	PGM	CVPR'23	32.2	30.9	29.5	30.1
	GUR	ICCV'23	17.0	17.7	12.9	14.4
	NG	ACM MM'24	27.4	25.0	26.1	25.4
	Ours	-	**41.8**	**40.1**	**41.4**	**40.1**

Table 1: Comparison of Rank-1 and mAP (%) on BUPTCampus. 'CD' denotes the Cross Domain baseline, where models are pre-trained on the source domain and directly tested on the target domain without fine-tuning.

	Method	Reference	RGB→IR		IR→RGB	
			Rank-1	mAP	Rank-1	mAP
Supervised	MITML	CVPR'22	63.7	45.3	64.5	47.7
	IBAN	TCSVT'23	65.0	48.8	69.6	51.0
	SADSTRM	Arxiv'23	65.3	49.5	67.7	51.8
	SAADG	ACM MM'23	69.2	53.8	73.1	56.1
	CST	TMM'24	69.4	51.2	72.6	53.0
	AuxNet	TIFS'24	51.1	46.0	54.6	48.7
Unsupervised	CD	-	40.1	26.6	42.8	24.8
	ADCA	ACM MM'22	35.7	23.8	33.0	23.2
	PGM	CVPR'23	36.8	23.9	33.6	23.4
	GUR	ICCV'23	9.9	5.9	8.9	5.6
	NG	ACM MM'24	24.5	12.0	22.8	11.9
	Ours	-	**63.8**	**50.9**	**64.3**	**49.3**

Table 2: Comparison of Rank-1 and mAP (%) on HITSZ-VCM. 'CD' denotes the Cross Domain baseline, where models are pre-trained on the source domain and directly tested on the target domain without fine-tuning.

achieves competitive or superior results compared to recent supervised methods, demonstrating its strong cross-modal matching capability under unsupervised settings.

Results on BUPTCampus On BUPTCampus, DSC achieves 41.8% Rank-1 and 40.0% mAP in the RGB→IR direction, outperforming the Cross-Domain baseline by 16.6% and 14.8%, respectively. In the IR→RGB direction, it reaches 41.3% Rank-1 and 40.1% mAP. Compared to supervised methods, DSC surpasses AlignGAN (28.0%) and LbA (32.1%), and closes the gap with DART (52.4%) and DEEN (53.7%). It also shows competitive performance against TF-CLIP (49.4%) in the IR→RGB setting.

Results on HITSZ-VCM On HITSZ-VCM, DSC obtains 63.8% Rank-1 and 50.9% mAP (RGB→IR), and 64.3% Rank-1 and 49.3% mAP (IR→RGB), outperforming the Cross-Domain baseline by over 20%. It outperforms supervised methods like MITML (63.7%) and IBAN (65.0%), and approaches the performance of SADSTRM (65.3%) and SAADG (69.2%) in RGB→IR. In IR→RGB, DSC performs on par with IBAN (69.6%) and ACST (72.6%), demonstrating strong effectiveness without labeled data.

Component				RGB→IR		IR→RGB	
B	Clu.	DSJL	DSLU	R@1	mAP	R@1	mAP
✓	✗	✗	✗	25.2	26.8	19.1	19.8
✓	✓	✗	✗	25.3	27.5	25.4	25.7
✓	✓	✓	✗	26.5	27.7	26.6	27.1
✓	✓	✗	✓	38.8	39.3	39.2	38.1
✓	✓	✓	✓	**41.8**	**40.0**	**41.3**	**40.1**

Table 3: Ablation study on BUPTCampus dataset showing the impact of different components on mAP and CMC (%). "R@1" denotes Rank-1.

DSJL Component		RGB→IR		IR→RGB	
Intra	Inter	R@1	mAP	R@1	mAP
✗	✗	38.8	39.3	39.2	38.1
✗	✓	38.2	38.0	36.5	37.4
✓	✗	39.4	37.3	37.9	37.7
✓	✓	**41.8**	**40.0**	**41.3**	**40.1**

Table 4: Ablation study on the internal components of DSJL on the BUPTCampus dataset. "Intra" and "Inter" denote intra- and inter-modality neighbor learning, respectively.

Ablation Study

Analysis of DSLU and its Components As shown in Tab. 3, adding DSLU to the baseline with clustering (**B+Clu.**) significantly improves performance by refining pseudo-labels through dynamic-static consistency. To further assess its design, we ablate DSLU's two components in Tab. 5. Inter-modal unification provides notable gains by enhancing cross-modal label consistency, while intra-modal correction alone yields limited improvements. Their combination achieves the best performance (41.8% Rank-1 / 40.0% mAP for RGB→IR and 41.3% / 40.1% for IR→RGB), confirming their complementary roles in robust pseudo-label learning.

Analysis of DSJL and its Components Based on B+Clu.+DSLU, introducing DSJL yields further gains by enhancing feature discrimination and modality alignment, as shown in Tab. 3. To further assess its design, we ablate DSJL's two components in Tab. 4, where removing both intra- and inter-modal neighbor learning results in significant drops, underscoring their importance. Intra-modal learning enhances local consistency, while inter-modal learning alone is less stable due to noisy cross-modal labels. Their combination achieves the best results (41.8% Rank-1 / 40.0% mAP for RGB→IR and 41.3% / 40.1% for IR→RGB), demonstrating their complementary strengths.

Retrieval Results Analysis Figure 3 compares retrieval results of the baseline and DSC on the HITSZ-VCM dataset. Query1 is a heavily occluded visible sequence; Query2

DSLU Component		RGB→IR		IR→RGB	
Intra	Inter	R@1	mAP	R@1	mAP
✗	✗	26.5	27.7	26.6	27.1
✗	✓	35.7	35.5	37.5	36.7
✓	✗	27.9	29.4	25.2	25.8
✓	✓	**41.8**	**40.0**	**41.3**	**40.1**

Table 5: Ablation study on the internal components of DSLU on the BUPTCampus dataset. "Intra" and "Inter" denote intra-modality correction and inter-modality unification.

λ_{dy}	λ_{st}	RGB→IR		IR→RGB	
		R@1	mAP	R@1	mAP
1.0	0.0	36.5	35.9	38.3	38.0
0.9	0.1	38.4	37.3	39.6	38.0
0.8	0.2	36.3	35.9	39.8	38.8
0.7	0.3	39.6	38.6	39.0	37.9
0.6	0.4	37.5	37.3	39.0	37.9
1.0	1.0	**41.8**	**40.0**	**41.3**	**40.1**

Table 6: Ablation study on intra-modal correction with different dynamic/static weight combinations on the BUPTCampus dataset. "R@1" denotes Rank-1.

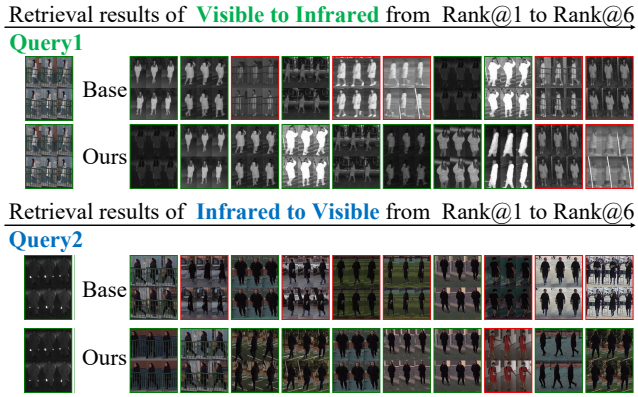

Figure 3: Visualization of retrieval results for IR→RGB and RGB→IR tasks,showing results from Rank@1 to Rank@6 on the HITSZ-VCM datasets.

is a low-resolution infrared sequence. For RGB→IR retrieval, the baseline struggles with occlusion, while DSC retrieves correct identities. For IR→RGB, resolution degradation misleads the baseline, whereas DSC maintains robust retrieval, demonstrating improved resilience to occlusion and resolution challenges.

Analysis of Crucial Parameters We analyze key parameters in DSJL and DSLU. As shown in Fig. 4(a), $\alpha = 0.9$ best balances neighbor diversity and reliability. Fig. 4(b) shows that top-$k = 20$ achieves optimal pseudo-label quality.

Table 6 shows that balanced similarity weights ($\lambda_{dy} = \lambda_{st} = 1.0$) yield the best performance (R@1 41.8%, mAP 40.0%). Imbalanced settings degrade results, confirming the necessity of leveraging both motion and appearance cues.

Analysis of Identity Distribution Figure 5 visualizes the feature distance distributions from BUPTCampus

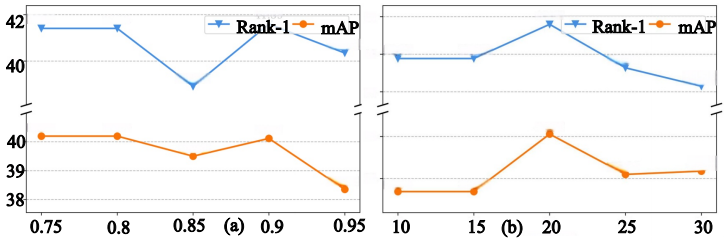

Figure 4: Impact of neighborhood threshold α in DSJL and top-k in DSLU on mAP performance of the proposed DSC framework on the BUPTCampus dataset.

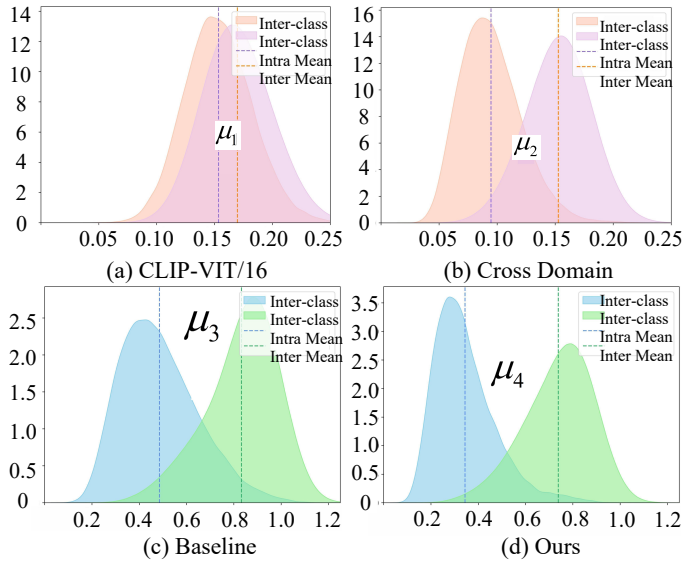

Figure 5: Density plots of feature distances for (a) CLIP-ViT/16, (b) Cross Domain, (c) Baseline, and (d) Our method, showing intra-class and inter-class distributions with mean distances (μ1 to μ4) indicated.

(source) to HITSZ-VCM (target) across four settings: (a) CLIP–ViT/16, (b) direct evaluation, (c) baseline UDA, and (d) our DIS framework. Compared to minimal separation in (a)–(c), DIS (d) achieves the largest inter-/intra-class separation μ_4, satisfying $\mu_1 < \mu_2 < \mu_3 < \mu_4$, by effectively compacting intra-class distances and expanding inter-class gaps.

Conclusion

We propose the Dynamic-Static Collaboration (DSC) framework for unsupervised visible-infrared video person re-identification (UDA-VVI-ReID). By jointly modeling dynamic motion cues and static appearance features through DSLU and DSJL, DSC mitigates modality gaps and enhances pseudo-label reliability. Experiments on HITSZ-VCM and BUPTCampus demonstrate its robustness under occlusion and low illumination, validating DSC's effectiveness in dynamic-static integration for scalable cross-modal video ReID.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant No.2022YFA1004100, in part by the Science and Technology Innovation Key R&D Program of Chongqing under Grant No. CSTB2023TIAD-STX0016, in part by the National Natural Science Foundation of China under Grants No. 62472060 and 62221005, in part by the Natural Science Foundation of Chongqing under Grants No. CSTB2024NSCQ-QCXMX0060 and CSTB2023NSCQ-LZX0061, in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant No. KJZD-K202300604.

References

Chen, C.; Ye, M.; Qi, M.; Wu, J.; Liu, Y.; and Jiang, J. 2022. Saliency and granularity: Discovering temporal coherence for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9): 6100–6112.

Cheng, D.; He, L.; Wang, N.; Zhang, S.; Wang, Z.; and Gao, X. 2023a. Efficient bilateral cross-modality cluster matching for unsupervised visible-infrared person reid. In *Proceedings of the 31st ACM international conference on multimedia*, 1325–1333.

Cheng, D.; Huang, X.; Wang, N.; He, L.; Li, Z.; and Gao, X. 2023b. Unsupervised visible-infrared person reid by collaborative learning with neighbor-guided label refinement. In *Proceedings of the 31st ACM international conference on multimedia*, 7085–7093.

Cheng, H.; Liu, M.-H.; Guo, Y.; Wang, T.; Nie, L.; and Kankanhalli, M. 2025. Fair Deepfake Detectors Can Generalize. In *NeurIPS*.

Cui, Z.; Zhou, J.; and Peng, Y. 2024. Dma: Dual modality-aware alignment for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 19: 2696–2708.

Dai, Z.; Wang, G.; Yuan, W.; Zhu, S.; and Tan, P. 2022. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian conference on computer vision*, 1142–1160.

Du, Y.; Lei, C.; Zhao, Z.; Dong, Y.; and Su, F. 2023. Video-based visible-infrared person re-identification with auxiliary samples. *IEEE Transactions on Information Forensics and Security*, 19: 1313–1325.

Feng, Y.; Chen, F.; Yu, J.; Ji, Y.; Wu, F.; Liu, T.; Liu, S.; Jing, X.-Y.; and Luo, J. 2024. Cross-Modality Spatial-Temporal Transformer for Video-Based Visible-Infrared Person Re-Identification. *IEEE Transactions on Multimedia*.

Ge, Y.; Chen, D.; and Li, H. 2020. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*.

Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; et al. 2020. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in neural information processing systems*, 33: 11309–11321.

Gong, Y.; Huang, L.; and Chen, L. 2022. Person re-identification method based on color attack and joint defence. In *CVPR, 2022*, 4313–4322.

Gong, Y.; Zhong, Z.; Qu, Y.; Luo, Z.; Ji, R.; and Jiang, M. 2024. Cross-modality perturbation synergy attack for person re-identification. *Advances in Neural Information Processing Systems*, 37: 23352–23377.

Li, H.; Liu, M.; Hu, Z.; Nie, F.; and Yu, Z. 2023a. Intermediary-guided Bidirectional Spatial-Temporal Aggregation Network for Video-based Visible-Infrared Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, H.; Xu, L.; Zhang, Y.; Tao, D.; and Yu, Z. 2023b. Adversarial Self-Attack Defense and Spatial-Temporal Relation Mining for Visible-Infrared Video Person Re-Identification. *arXiv preprint arXiv:2307.03903*.

Li, S.; Leng, J.; Kuang, C.; Tan, M.; and Gao, X. 2025. Video-Level Language-Driven Video-Based Visible-Infrared Person Re-Identification. *IEEE Transactions on Information Forensics and Security*.

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.

Lin, X.; Li, J.; Ma, Z.; Li, H.; Li, S.; Xu, K.; Lu, G.; and Zhang, D. 2022. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20973–20982.

Liu, M.; Cheng, H.; Wang, T.; Luo, X.; and Xu, X. 2025. Learning Real Facial Concepts for Independent Deepfake Detection. In *IJCAI*, 1585–1593.

Pang, Z.; Wang, C.; Zhao, L.; Liu, Y.; and Sharma, G. 2023. Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification. *IEEE Transactions on circuits and systems for video technology*, 34(4): 2706–2718.

Park, H.; Lee, S.; Lee, J.; and Ham, B. 2021. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12046–12055.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, 17–35. Springer.

Shi, J.; Zhang, Y.; Yin, X.; Xie, Y.; Zhang, Z.; Fan, J.; Shi, Z.; and Qu, Y. 2023. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11218–11228.

Wang, G.; Zhang, T.; Cheng, J.; Liu, S.; Yang, Y.; and Hou, Z. 2019a. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3623–3632.

Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.-Y.; and Satoh, S. 2019b. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 618–626.

Wei, Z.; Yang, X.; Wang, N.; and Gao, X. 2021. Syncretic modality collaborative learning for visible infrared person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 225–234.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, 5380–5389.

Wu, Z.; and Ye, M. 2023. Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9548–9558.

Yang, B.; Chen, J.; and Ye, M. 2023. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11069–11079.

Yang, B.; Chen, J.; and Ye, M. 2024. Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16870–16879.

Yang, B.; Ye, M.; Chen, J.; and Wu, Z. 2022a. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2843–2851.

Yang, M.; Huang, Z.; Hu, P.; Li, T.; Lv, J.; and Peng, X. 2022b. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14308–14317.

Ye, M.; Chen, C.; Shen, J.; and Shao, L. 2021a. Dynamic tri-level relation mining with attentive graph for visible infrared re-identification. *IEEE Transactions on Information Forensics and Security*, 17: 386–398.

Ye, M.; Ruan, W.; Du, B.; and Shou, M. Z. 2021b. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13567–13576.

Ye, M.; Shen, J.; J. Crandall, D.; Shao, L.; and Luo, J. 2020. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, 229–247. Springer.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021c. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 44(6): 2872–2893.

Ye, M.; Shen, J.; and Shao, L. 2020. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security*, 16: 728–739.

Ye, M.; Wu, Z.; Chen, C.; and Du, B. 2023. Channel augmentation for visible-infrared re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4): 2299–2315.

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 *AAAI Conference on Artificial Intelligence, AAAI*, volume 38, 6764–6772.

Zhang, M.; Xiao, Y.; Xiong, F.; Li, S.; Cao, Z.; Fang, Z.; and Zhou, J. T. 2022a. Person re-identification with hierarchical discriminative spatial aggregation. *IEEE Transactions on Information Forensics and Security*, 17: 516–530.

Zhang, Q.; Lai, C.; Liu, J.; Huang, N.; and Han, J. 2022b. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7349–7358.

Zhang, Y.; and Wang, H. 2023. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2153–2162.

Zhang, Y.; Yan, Y.; Lu, Y.; and Wang, H. 2021. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM international Conference on Multimedia*, 788–796.

Zhou, C.; Li, J.; Li, H.; Lu, G.; Xu, Y.; and Zhang, M. 2023. Video-based visible-infrared person re-identification via style disturbance defense and dual interaction. In *Proceedings of the 31st ACM International Conference on Multimedia*, 46–55.