Learning Source-free Domain Adaptation for Visible-Infrared Person Re-Identification

Yongxiang Li[1], **Yanglin Feng**[1], **Yuan Sun**[2], **Dezhong Peng**[1,3], **Xi Peng**[1], **Peng Hu**[1]*
[1]College of Computer Science, Sichuan University, Chengdu, China.
[2]National Key Laboratory of Fundamental Algorithms and Models
for Engineering Numerical Simulation, Sichuan University, Chengdu, China.
[3]Tianfu Jincheng Laboratory, Chengdu, China.
{rhythmli.scu, penghu.ml}@gmail.com

Abstract

In this paper, we investigate source-free domain adaptation (SFDA) for visible-infrared person re-identification (VI-ReID), aiming to adapt a pre-trained source model to an unlabeled target domain without access to source data. To address this challenging setting, we propose a novel learning paradigm, termed Source-Free Visible-Infrared Person Re-Identification (SVIP), which fully exploits the prior knowledge embedded in the source model to guide target domain adaptation. The proposed framework comprises three key components specifically designed for the source-free scenario: 1) a Source-Guided Contrastive Learning (SGCL) module, which leverages the discriminative feature space of the frozen source model as a reference to perform contrastive learning on the unlabeled target data, thereby preserving discrimination without requiring source samples; 2) a Residual Transfer Learning (RTL) module, which learns residual mappings to adapt the target model's representations while maintaining the knowledge from the source model; and 3) a Structural Consistency-Guided Cross-modal Alignment (SCCA) module, which enforces reciprocal structural constraints between visible and infrared modalities to identify reliable cross-modal pairs and achieve robust modality alignment without source supervision. Extensive experiments on benchmark datasets demonstrate that SVIP substantially enhances target domain performance and outperforms existing unsupervised VI-ReID methods under source-free settings. Code is available at https://github.com/LYXRhythm/SVIP.

1 Introduction

Visible-Infrared Person Re-Identification (VI-ReID) aims to match pedestrian images captured under visible and infrared modalities, achieving consistent cross-modality correspondence across varying lighting conditions, which is particularly valuable in various applications such as nighttime security and smart surveillance [1, 2, 3, 4, 5, 6, 7]. By leveraging complementary visual cues from multiple sensors, VI-ReID systems can maintain consistent performance in both daytime and nighttime environments. However, the majority of existing VI-ReID methods rely heavily on large-scale, well-annotated datasets with aligned visible and infrared images. Constructing such datasets demands extensive human annotation and careful sensor calibration, which is labor-intensive, time-consuming, and increasingly constrained by privacy regulations. These limitations severely hinder the scalability and real-world deployment of VI-ReID systems.

To reduce dependence on labeled data in the target domain, Unsupervised Domain Adaptation (UDA) has emerged as a promising alternative. UDA methods aim to transfer knowledge from a labeled

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-0690

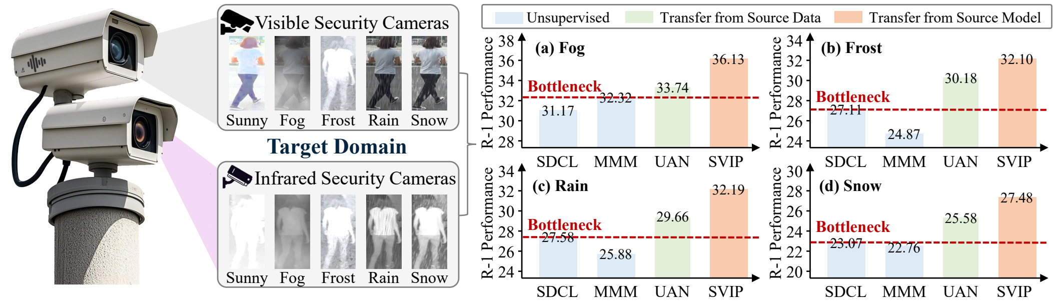


Figure 1: Our Observations. Despite recent advances, state-of-the-art unsupervised methods such as SDCL [11] and MMM [12] often suffer from performance limitations due to insufficient supervision in the target domain. Incorporating source-domain knowledge through labeled data (for example, UAN [13]) or pre-trained models (such as our proposed SVIP) can significantly enhance target-domain representation and adaptation. For example, we conduct experiments using SYSU-MM01 [14] as the source domain and LLCM-W under (a) *Fog*, (b) *Frost*, (c) *Rain*, (d) *Snow* conditions as target domains (see Section 4.1) to validate this observation.

source domain to an unlabeled target domain [8, 9, 10], achieving notable improvements over purely unsupervised methods, as illustrated in Figure 1. Despite their effectiveness, conventional UDA approaches typically assume joint access to source and target data during training. This assumption is often unrealistic in practical VI-ReID applications, where source data may contain sensitive personal information that cannot be shared or retained due to legal and ethical constraints. Moreover, the need to repeatedly load and process large-scale source datasets introduces significant computational overhead, especially in edge computing scenarios or resource-constrained environments. These issues collectively restrict the deployment of conventional UDA pipelines in real-world systems.

To address these practical constraints, Source-Free Domain Adaptation (SFDA) has recently garnered attention [15, 16, 17, 18]. SFDA eliminates the need for source data during adaptation by leveraging a pre-trained source model as the sole knowledge carrier. This setting naturally supports privacy-preserving and lightweight model adaptation, aligning more closely with real-world requirements. However, existing SFDA methods are predominantly designed for unimodal tasks and do not generalize well to VI-ReID, where domain shifts and modality gaps must be addressed simultaneously. In the absence of paired visible-infrared samples or annotated target identities, cross-modal association becomes highly uncertain, often leading to noisy feature learning and suboptimal adaptation. These challenges necessitate a new SFDA framework tailored to the unique characteristics of VI-ReID, capable of learning modality-invariant and identity-discriminative representations without relying on source data or cross-modal supervision.

Motivated by these observations, we propose a novel learning paradigm, termed **S**ource-Free Domain Adaptation for **V**isible-**I**nfrared **P**erson Re-Identification (SVIP). Specifically, our contributions include: 1) the Source Guided Contrastive Learning (SGCL) mechanism, which improves supervision reliability by integrating dual clustering results from both source and target models on target domain samples. It adaptively fuses supervisory signals weighted by confidence scores to mitigate category inconsistency and label noise across modalities, enhancing the discriminative power of the target model; 2) the Residual Transfer Learning (RTL) mechanism, which introduces a feature residual distillation loss to explicitly align intermediate feature representations of source and target models under identical inputs, preserving the core discriminative structure from the source domain, thus improving structural stability and generalization; 3) the Structural Consistency Guided Cross-modal Alignment (SCCA) mechanism, which exploits the source model's structural semantic knowledge derived from paired visible-infrared images to guide the target model in mining potential cross-modal similar samples within the unpaired target domain, thereby achieving stable and reliable cross-modal feature alignment and further enhancing consistency and robustness of cross-modal representations. Collectively, SVIP effectively transfers discriminative and structural knowledge from the source model to overcome key challenges in the unsupervised target domain, including knowledge transfer difficulty, and insufficient cross-modal alignment, significantly boosting the adaptability and performance of the target model in cross-modal person re-identification.

The main contributions of this work could be summarized as follows:

https://doi.org/10.52202/085713-0690

- We propose a novel SFDA framework, SVIP, for VI-ReID that adapts a pre-trained source model to an unlabeled target domain without requiring access to any source data. To the best of our knowledge, this is among the first works exploring SFDA specifically for VI-ReID.
- SVIP incorporates three key mechanisms: Source Guided Contrastive Learning (SGCL) to enhance supervision reliability via confidence-weighted dual clustering; Residual Transfer Learning (RTL) to align intermediate features and preserve source discriminative structures; and Structural Consistency Guided Cross-modal Alignment (SCCA) to exploit source structural knowledge for robust cross-modal feature alignment in the unpaired target domain.
- Extensive experiments under diverse domain adaptation settings demonstrate that SVIP consistently outperforms nine state-of-the-art baselines, validating its effectiveness and superiority.

2 Related Works

2.1 Source-free Domain Adaptation

Source-free domain adaptation (SFDA) has emerged as a crucial paradigm that addresses the limitations of conventional unsupervised domain adaptation (UDA) by eliminating the requirement to access source domain data during the adaptation process [19, 16, 15, 17, 13]. This characteristic makes SFDA particularly well-suited to practical scenarios that impose strict privacy constraints, involve decentralized data storage, or demand efficient model deployment in distributed computing environments. Existing SFDA methods can generally be divided into two broad categories. 1) Generation-based methods aim to learn domain-invariant representations by synthesizing or transforming target domain samples to approximate the distribution of source domain data. This enables the application of conventional UDA frameworks to reduce domain shifts [20, 18, 21, 17, 22]. 2) Self-training methods adopt an iterative learning strategy in which pseudo-labels generated from model predictions on unlabeled target data serve as supervision signals to progressively refine the model in a fully unsupervised manner [23, 24, 25, 26, 27]. While these approaches have demonstrated effectiveness in unimodal settings, their extension to scenarios involving cross-modal or multi-modal data remains limited. Moreover, they often fall short in fully capturing the rich semantic correlations and intrinsic structure present in heterogeneous target domains, which is a significant challenge in complex tasks such as re-identification.

2.2 Domain Adaptation for Person ReID

Person Re-Identification (ReID) faces significant challenges due to large domain shifts caused by variations in environment, camera styles, and modalities [28, 29, 30]. Unsupervised Domain Adaptation (UDA) has attracted much attention for its ability to adapt models to unlabeled target domains without manual annotation [31, 32, 33]. However, UDA methods typically require access to labeled source data during adaptation, which incurs high computational costs and raises privacy concerns, especially in ReID scenarios subject to strict data regulations. These limitations hinder the practical deployment of UDA models. SFDA addresses these issues by adapting pretrained source models to target domains without accessing source data [34, 35, 36]. SFDA aligns well with privacy preservation and deployment constraints. Nonetheless, existing SFDA methods for ReID are primarily designed for unimodal settings and often fail to exploit the intrinsic structure and semantic relationships of the target domain [36, 34], limiting their adaptation effectiveness. In addition, the absence of identity overlap between source and target domains further complicates feature transfer and undermines pseudo-label reliability. These limitations pose significant challenges in mitigating label noise, narrowing modality discrepancies, and achieving effective cross-modal alignment, all of which are crucial for advancing SFDA in ReID.

3 Methodology

3.1 Problem Statement and Notations

For clarity, we define the key notations and terminology used in this paper. Let Φ_S denote the source model trained on a labeled visible-infrared person re-identification (VI-ReID) dataset D_S. Additionally, let the unlabeled and unpaired VI-ReID dataset in the target domain be denoted as

 https://doi.org/10.52202/085713-0690

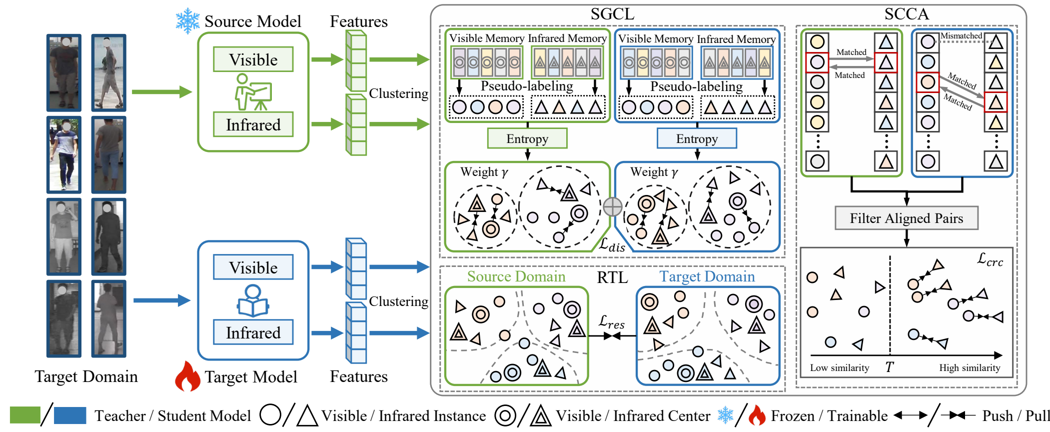


Figure 2: Overview of the proposed SVIP framework. SVIP integrates three core mechanisms: 1) Source Guided Contrastive Learning (SGCL) improves supervision reliability by combining dual clustering from source and target models with confidence-weighted supervision; 2) Residual Transfer Learning (RTL) aligns intermediate features between source and target models via residual distillation to preserve discriminative structures and enhance generalization; 3) Structural Consistency Guided Cross-modal Alignment (SCCA) leverages source domain structural knowledge to guide stable cross-modal feature alignment in the unpaired target domain. Together, these components enable effective source-free domain adaptation for VI-ReID.

$D_T = \left\{\{\boldsymbol{x}_i^{\mathcal{V}}\}_{i=1}^{N^{\mathcal{V}}}, \{\boldsymbol{x}_j^{\mathcal{I}}\}_{j=1}^{N^{\mathcal{I}}}\right\}$, where $N^{\mathcal{V}}$ and $N^{\mathcal{I}}$ are the numbers of visible images $\boldsymbol{x}_i^{\mathcal{V}}$ and infrared images $\boldsymbol{x}_j^{\mathcal{I}}$, respectively. For convenience, we denote D_T more generally as $D_T = \left\{\{\boldsymbol{x}_i^{\mathcal{P}}\}_{i=1}^{N^{\mathcal{P}}}\right\}$, where $\mathcal{P} \in \{\mathcal{V}, \mathcal{I}\}$. Due to the domain shift between the source and target domains, the performance of Φ_S can degrade significantly in practical scenarios, such as those involving environmental changes (e.g., variations in weather conditions). Our objective is to adapt Φ_S to the target domain D_T and obtain a target-specific model Φ_T, without accessing any source domain samples or ground-truth labels in the target domain. An overview of the SFDA framework for VI-ReID is illustrated in Figure 2. The main motivation of our method lies in the observation that the source model Φ_S contains rich discriminative knowledge and cross-modal alignment capabilities. Selectively transferring this knowledge to the target domain can enhance the target model Φ_T in terms of both discrimination and cross-modal semantic understanding, thereby overcoming the limitations of existing unsupervised methods.

3.2 Source Guided Contrastive Learning

Existing unsupervised methods typically rely on clustering to generate pseudo-labels, which provide pseudo-supervision signals for training samples. However, due to the absence of prior knowledge, the clustering results are often susceptible to modality gaps, leading to a degradation in pseudo-label quality and resulting in label noise issues [7, 37, 38, 39, 40, 41]. To alleviate this problem, we propose a Source Guided Contrastive Learning (SGCL) mechanism that leverages the prior knowledge embedded in the source model Φ_S to guide the target model Φ_T towards more reliable supervision signals.

Specifically, SGCL adaptively integrates the dual predictions of target domain samples from both the source model and the target model based on their confidence scores. This produces a more trustworthy supervisory signal, enhancing the understanding and adaptability of the target model Φ_T to the target domain D_T. First, we extract features of target domain samples using the source model Φ_S and the target model Φ_T, and apply DBSCAN clustering to obtain two sets of cluster centers, denoted as C_S and C_T:

$$C_S = \text{DBSCAN}\big(\Phi_S(\boldsymbol{x}_i^{\mathcal{P}})\big), \quad C_T = \text{DBSCAN}\big(\Phi_T(\boldsymbol{x}_i^{\mathcal{P}})\big). \tag{1}$$

To enhance the stability of clustering representations, we introduce two memory banks: $M_S = \{\boldsymbol{m}_k^S\}_{k=1}^{K_S}$ and $M_T = \{\boldsymbol{m}_k^T\}_{k=1}^{K_T}$, which store the cluster center features obtained by the source and target models, respectively, in the current training cycle. Here, K_S and K_T denote the number of clusters for the source and target models, respectively. The memory banks are updated in each training cycle using the exponential moving average (EMA) strategy:

$$\boldsymbol{m}_k^S \leftarrow \eta \boldsymbol{m}_k^S + (1-\eta) \cdot \frac{1}{|C_k^S|} \sum_{i \in C_k^S} \Phi_S(\boldsymbol{x}_i^{\mathcal{P}}), \tag{2}$$

$$\boldsymbol{m}_k^T \leftarrow \eta \boldsymbol{m}_k^T + (1-\eta) \cdot \frac{1}{|C_k^T|} \sum_{i \in C_k^T} \Phi_T(\boldsymbol{x}_i^{\mathcal{P}}), \tag{3}$$

where $\eta \in [0, 1)$ is the momentum coefficient, and C_k^S and C_k^T represent the sample index sets of the k-th cluster for the source and target models, respectively.

Given a sample $\boldsymbol{x}_i^P$, we compute its similarity to all cluster centers from both models and obtain a probability distribution over clusters using the Softmax function:

$$p_S(k) = \frac{\exp(\langle \Phi_S(x_i^P), m_k^S \rangle / \tau)}{\sum_{j=1}^{K_S} \exp(\langle \Phi_S(x_i^P), m_j^S \rangle / \tau)}, \tag{4}$$

$$p_T(k) = \frac{\exp(\langle \Phi_T(x_i^P), m_k^T \rangle / \tau)}{\sum_{j=1}^{K_T} \exp(\langle \Phi_T(x_i^P), m_j^T \rangle / \tau)}, \tag{5}$$

where τ is a temperature parameter that controls the smoothness of the distribution. Then, the prediction confidence is estimated based on the entropy of the distributions, and a fusion weight is computed accordingly:

$$\gamma = \frac{\exp(-H(p_S))}{\exp(-H(p_S)) + \exp(-H(p_T))}, \tag{6}$$

where $\boldsymbol{p}_S = [p_S(1), \cdots, p_S(k), \cdots, p_S(K_S)]$ and $\boldsymbol{p}_T = [p_T(1), \cdots, p_T(k), \cdots, p_T(K_T)]$. $H(\cdot)$ denotes the Shannon entropy.

Considering the possible inconsistency in the number of clusters between the source and target models (i.e., $K_S \neq K_T$), directly fusing the full Softmax distributions would lead to dimensional mismatch issues. Therefore, instead of fusing the distributions themselves, we fuse the features based on the respective assigned cluster centers of the sample in each model. This design avoids dimensional inconsistency and enhances fusion accuracy. In other words, for a given sample $\boldsymbol{x}_i^P$, the optimization objective is defined as:

$$\mathcal{L}_{\text{dis}} = \frac{1}{N_b} \sum_{\mathcal{P} \in \{\mathcal{V}, \mathcal{I}\}} \sum_{i=1}^{N_b} \gamma \cdot I(\Phi_S(\boldsymbol{x}_i^{\mathcal{P}}), m_+^S) + (1-\gamma) \cdot I(\Phi_T(\boldsymbol{x}_i^{\mathcal{P}}), \boldsymbol{m}_+^T), \tag{7}$$

where m_+^S and m_+^T denote the assigned cluster centers for the sample in the source and target domains, respectively. $I(\cdot, \cdot)$ represents the InfoNCE loss [42], and N_b is the batch size.

This fusion strategy fully exploits the multi-view clustering structures of the same sample from both the source and target models. By adaptively balancing their supervision signals based on prediction confidence, it enables the target model to benefit from the prior knowledge embedded in the source model. This, in turn, mitigates the clustering noise caused by modality discrepancies and enhances the generalization and adaptability of the target model in the target domain.

3.3 Residual Transfer Learning

Although the source-guided contrastive mechanism introduced in the previous section provides stable supervision for the target model to adapt to the target domain, it remains limited in maintaining the consistency of discriminative behavior. To address this issue, we propose a Residual Transfer Learning (RTL) mechanism to explicitly align source and target domains. Inspired by the principle of preserving core discriminative knowledge in domain adaptation, this mechanism formalizes the discrepancy in discriminative behavior between the source model and the target model under the same input as a form of feature distillation constraint. This constraint encourages the target model to continuously align its intermediate feature representations with those of the source model during the adaptation process, thereby preserving structural knowledge and maintaining consistency in prediction behavior.

 https://doi.org/10.52202/085713-0690

Specifically, given a target-domain sample $\boldsymbol{x}_i^{\mathcal{P}}$, the residual distillation loss is defined as:

$$\mathcal{L}_{\text{res}} = \left\| \Phi_S(\boldsymbol{x}_i^{\mathcal{P}}) - \Phi_T(\boldsymbol{x}_i^{\mathcal{P}}) \right\|_2^2 . \tag{8}$$

$\mathcal{L}_{\text{res}}$ guides the target model to retain the structural knowledge of the source model in the discriminative feature space. Compared to probability distribution alignment approaches, this method does not rely on consistency in the output space of the two models, making it more suitable for the scenarios with label space shifts or mismatched output dimensions. Therefore, it offers greater flexibility and stability. Beyond the SGCL, the RTL mechanism further strengthens the continuous guidance from the source model to the target model, thus achieving a more effective balance between generalization and adaptation capabilities.

3.4 Structural Consistency Guided Cross-modal Alignment

In an unlabeled and unpaired target domain, the absence of explicit cross-modal correspondences poses significant challenges to robust alignment. To alleviate this issue, we propose a Structural Consistency Guided Cross-modal Alignment (SCCA) mechanism, which leverages stable and reliable structural semantic knowledge from the source model to assist the target model in discovering potential semantically consistent image pairs, thereby facilitating consistent cross-modal representation learning.

The source model Φ_S, trained with supervision on paired visible-infrared images, has acquired strong capabilities in modeling cross-modal structural consistency. To transfer this capability to the unpaired target domain, we jointly exploit the feature modeling of both the source model and the current target model to estimate potential cross-modal similarities. Given the unpaired target-domain samples $\{\boldsymbol{x}_i^{\mathcal{V}}\}_{i=1}^{N^{\mathcal{V}}}$ and $\{\boldsymbol{x}_j^{\mathcal{I}}\}_{j=1}^{N^{\mathcal{I}}}$, the structural similarity between a visible image $\boldsymbol{x}_i^{\mathcal{V}}$ and an infrared image $\boldsymbol{x}_j^{\mathcal{I}}$ is defined as:

$$S_{ij} = \frac{1}{2} \left(\langle \Phi_S(\boldsymbol{x}_i^{\mathcal{V}}), \Phi_S(\boldsymbol{x}_j^{\mathcal{I}}) \rangle + \langle \Phi_T(\boldsymbol{x}_i^{\mathcal{V}}), \Phi_T(\boldsymbol{x}_j^{\mathcal{I}}) \rangle \right), \tag{9}$$

where Φ_S provides structural prior guidance, and Φ_T reflects the adaptation status of the target model. The similarity score S_{ij} integrates source-domain knowledge and target-domain modeling capacity, and is used to construct cross-modal similarity vectors $\boldsymbol{S}_i^{\mathcal{V}\to\mathcal{I}} = [S_{ij} | S_{i1}, \cdots, S_{iN^{\mathcal{I}}}]$ and $\boldsymbol{S}_i^{\mathcal{I}\to\mathcal{V}} = [S_{ji} | S_{j1}, \cdots, S_{jN^{\mathcal{V}}}]$, which are employed to estimate pseudo matching pairs.

To enhance the accuracy and robustness of the pseudo matches, two structural constraints are imposed: 1) Reciprocity Constraint ($\mathcal{C}_1$): Each image pair must be the most similar to each other, *i.e.*, , Equation (10). 2) Lower-bound Similarity Constraint ($\mathcal{C}_2$): The structural similarity between the image pair must exceed a predefined threshold T, *i.e.*, , Equation (11).

These constraints are formally defined as:

$$\mathcal{C}_1 = \begin{cases} \underset{i\in\{1,\cdots,N^{\mathcal{V}}\}}{\text{arg max}} \; S_{\underset{j\in\{1,\cdots,N^{\mathcal{I}}\}}{\text{arg max}}\, S_{ij}\, i} = i \\ \underset{j\in\{1,\cdots,N^{\mathcal{I}}\}}{\text{arg max}} \; S_{\underset{i\in\{1,\cdots,N^{\mathcal{V}}\}}{\text{arg max}}\, S_{ji}\, j} = j \end{cases}, \tag{10}$$

$$\mathcal{C}_2 = (S_{ij} \geq T) \wedge (S_{ji} \geq T) . \tag{11}$$

An image pair $(\boldsymbol{x}_i^{\mathcal{V}}, \boldsymbol{x}_j^{\mathcal{I}})$ is regarded as a reliable pseudo match if both $\mathcal{C}_1$ and $\mathcal{C}_2$ are satisfied. Such pairs are marked using a binary indicator $\mathcal{A}_{ij} = \mathbb{1}[(\boldsymbol{x}_i^{\mathcal{V}}, \boldsymbol{x}_j^{\mathcal{I}}) \models \mathcal{C}_1 \wedge \mathcal{C}_2]$. Under the guidance of the structural priors from the source model, this strategy significantly mitigates the uncertainty caused by the lack of pairing.

On this basis, to further improve the compactness of cross-modal pseudo matches in the feature space, we introduce a Cross-modal Reciprocal Consistency Loss:

$$\mathcal{L}_{crc} = -\frac{1}{N_b} \sum_{\mathcal{P}\in\{\mathcal{V},\mathcal{I}\}} \sum_{i=1}^{N_b} \mathcal{A}_{ij} \log \frac{\exp(\langle \Phi_T(\boldsymbol{x}_i^{\mathcal{P}}), \Phi_T(\boldsymbol{x}_j^{\mathcal{Q}}) \rangle / \tau)}{\sum_{t=1}^{N^{\mathcal{Q}}} \exp(\langle \Phi_T(\boldsymbol{x}_i^{\mathcal{P}}), \Phi_T(\boldsymbol{x}_j^{\mathcal{Q}}) \rangle / \tau)}, \quad \text{s.t. } \mathcal{Q} \neq \mathcal{P}, \tag{12}$$

where τ denotes a temperature parameter. This loss employs reliable pseudo matches as supervisory signals and utilizes contrastive learning to enhance semantic consistency between cross-modal images, thereby improving the alignment and discriminative power of cross-modal features.

3.5 Overall Training Procedure

Given that source data is unavailable, the proposed method optimizes the target model Φ_T in each iteration using only the unlabeled data from the target domain. The entire training process is performed without access to any source domain samples, relying solely on the prior knowledge embedded in the source model Φ_S to guide cross-modal semantic transfer. The optimization objective for the target model integrates three key components and is defined as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{dis} + \lambda_1 \mathcal{L}_{res} + \lambda_2 \mathcal{L}_{crc}, \tag{13}$$

where λ_1 and λ_2 are weighting coefficients.

4 Experiments

4.1 Experiment Settings

Datasets: To systematically investigate domain adaptation in VI-ReID, we conduct experiments on three widely used datasets: SYSU-MM01 [14], RegDB [43], and LLCM [44]. These datasets provide a comprehensive foundation for exploring the challenges and opportunities in cross-modal person re-identification across varying domains.

Domain Adaptation Settings: To comprehensively evaluate the adaptation performance across diverse domains, we define two domain adaptation settings: **1)** ***Basic Setting*****:** This setting is designed to evaluate the model's ability to handle common domain shifts, including variations in illumination, camera viewpoints, and background complexity. Four adaptation scenarios are considered: (i) SYSU-MM01 $\rightarrow$ RegDB, (ii) SYSU-MM01 $\rightarrow$ LLCM, (iii) LLCM $\rightarrow$ RegDB, and (iv) LLCM $\rightarrow$ SYSU-MM01. SYSU-MM01 and LLCM are used as source domains, while RegDB, LLCM, and SYSU-MM01 serve as target domains accordingly. Due to its relatively small scale, RegDB is not used as a source domain. **2)** ***Weather Setting*****:** To further evaluate model robustness under realistic environmental variations, we introduce five weather conditions to both RegDB and LLCM, including *Sunny*, *Fog*, *Frost*, *Rain*, and *Snow*. Each condition is simulated at three severity levels, resulting in corrupted versions of the datasets, denoted as RegDB-W and LLCM-W. In this setting, the model trained on SYSU-MM01 is adopted as the source model, and domain adaptation is performed on each corrupted target domain. The experiment follows the standard evaluations in VI-ReID [45, 46], including Cumulative Matching Characteristic (CMC), and mean Average Precision (mAP).

Implementation Details: During training, pedestrian images are resized to 288$\times$144 pixels. The AGW [47] network is employed as the feature extractor. The parameters of the source model remain fixed throughout training. The target model is optimized using the Adam optimizer with a weight decay of 8×10^{-4}. The initial learning rate is set to 5×10^{-4} and is reduced by a factor of 10 every 20 epochs. The batch size N_b is 128, and training proceeds for 50 epochs. The momentum parameter η is fixed at 0.1, while the temperature hyperparameter τ is set to 0.05. The clustering algorithm DBSCAN is applied with an epsilon value of 0.6 and a minimum sample size of 4. The threshold T is maintained at 0.5. The trade-off parameters λ_1 and λ_2 are further analyzed in the Supplementary Material. All experiments are performed on an Ubuntu 20.04 system equipped with four NVIDIA RTX 3090 GPUs.

Training of the Source Model: For the source model, our method could use the model trained by any VI-ReID method. Without loss of generality, we follow [23] and present a simple yet representative approach with the triplet loss function [45, 48] to train a source model for our experiments:

$$\mathcal{L}_{tri} = \sum_{\mathcal{P}\in\{\mathcal{V},\mathcal{I}\}} \sum_{i=1}^{N_b} \max\Big(\|\Phi_S(\boldsymbol{x}_i^a) - \Phi_S(\boldsymbol{x}_i^p)\|_2^2 - \|\Phi_S(\boldsymbol{x}_i^a) - \Phi_S(\boldsymbol{x}_i^n)\|_2^2 + \tau_{tri}, 0\Big), \tag{14}$$

where $\boldsymbol{x}_i^a \in D_S$ represents anchor in source training set. $\boldsymbol{x}_i^p \in D_S$ and $\boldsymbol{x}_i^n \in D_S$ are positive (same identity) and negative (different identity) samples of $\boldsymbol{x}_i^a$. τ_{tri} is the margin, which is set to 0.3.

https://doi.org/10.52202/085713-0690

Table 1: Comparisons with the state-of-the-art methods under *Basic Setting*. R-1 (%), mAP (%) and mINP (%) are reported. The best results are marked in **bold**, and the second-best results are underlined.

	Methods	SYSU-MM01 → LLCM				SYSU-MM01 → RegDB				LLCM → SYSU-MM01				LLCM → RegDB			
		V2I		I2V		V2T		T2V		All Search		Indoor Search		T2V		T2V	
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
	Source Only	12.47	12.91	12.21	13.13	31.26	25.71	28.95	25.92	17.89	20.35	20.91	20.56	24.51	23.29	25.48	21.14
UVI-ReID	OTAL [49]	17.88	20.46	14.97	18.66	32.90	29.70	32.10	28.60	29.90	27.10	29.80	38.80	32.90	29.70	32.10	28.60
	CCLNet [50]	30.33	27.12	30.11	26.67	69.94	65.53	70.17	66.66	54.03	50.19	56.68	65.12	69.94	65.53	70.17	66.66
	PGM [51]	30.95	33.68	37.82	43.52	69.48	65.41	69.85	65.17	57.27	51.78	56.23	62.74	69.48	65.41	69.85	65.17
	GUR [52]	31.47	34.77	29.68	33.38	73.91	70.23	75.00	69.94	60.95	56.99	64.22	69.49	73.91	70.23	75.00	69.94
	SDCL [11]	38.86	45.19	46.34	43.35	86.91	78.92	85.76	77.25	64.49	63.24	71.37	76.90	86.91	78.92	85.76	77.25
	MMM [12]	35.39	42.57	45.05	46.26	89.70	80.50	85.80	77.00	65.90	61.80	70.30	74.90	89.70	80.50	85.80	77.00
SFDA	CMT [53]	31.22	33.53	37.37	40.77	72.06	65.01	72.12	62.98	42.15	38.20	45.33	41.87	50.62	44.35	53.21	47.89
	LEAD [15]	29.37	32.73	36.16	38.02	75.99	66.38	70.91	62.90	40.88	36.54	43.17	40.12	48.95	42.76	51.03	45.67
	DRU [54]	28.48	33.80	36.55	38.85	76.26	68.54	71.68	64.43	41.72	37.89	44.65	41.03	49.81	43.52	52.14	46.35
	SVIP (Ours)	**40.50**	**46.20**	**47.85**	**47.30**	**90.10**	**81.20**	**86.50**	**77.80**	**66.80**	**63.50**	**71.60**	**77.40**	**90.20**	**81.60**	**86.10**	**78.50**

Table 2: Comparisons with state-of-the-art methods under the *Weather Setting* (Severity Level 3), reporting R-1 (%) accuracy. The best results are marked in **bold**, and the second-best results are underlined.

	Methods	LLCM-W (Visible to Infrared)						RegDB-W (Visible to Thermal)					
		Sunny	Fog	Frost	Rain	Snow	*Avg.*	Sunny	Fog	Frost	Rain	Snow	*Avg.*
	Source Only	8.07	8.20	9.82	8.19	8.70	8.60	16.06	20.86	16.33	20.18	11.55	16.06
UVI-ReID	OTAL [49]	12.17	13.56	10.01	9.72	12.90	11.67	28.14	30.96	24.44	30.84	17.98	26.47
	CCLNet [50]	23.05	25.86	21.88	27.04	19.82	23.53	42.27	46.84	36.95	45.15	27.26	39.69
	PGM [51]	25.16	26.29	25.30	22.52	18.00	23.45	49.77	54.63	43.39	53.55	32.09	46.69
	GUR [52]	26.39	29.61	22.89	21.76	17.68	23.67	42.12	57.87	42.41	54.09	33.13	45.93
	SDCL [11]	26.93	31.17	27.11	27.58	23.07	27.17	50.87	61.08	49.21	64.26	39.01	52.89
	MMM [12]	25.63	32.32	24.87	25.88	22.76	26.29	49.17	64.28	47.57	64.04	39.68	52.95
SFDA	CMT [53]	23.24	29.74	20.49	21.16	15.97	22.12	48.03	60.36	43.63	62.01	32.78	49.36
	LEAD [15]	22.34	30.16	19.79	19.75	17.51	21.91	45.89	63.23	42.06	60.81	33.73	49.15
	DRU [54]	20.57	31.48	23.33	22.80	18.16	23.27	46.46	64.74	43.30	61.34	39.82	51.13
	SVIP (Ours)	**28.02**	**36.13**	**32.10**	**32.19**	**27.48**	**30.87**	**54.32**	**65.92**	**49.54**	**67.46**	**46.05**	**57.81**

4.2 Comparison with the State-of-the-Arts

We compare our SVIP method with nine state-of-the-art approaches, categorized into two groups. The first group comprises six Unsupervised Visible-Infrared Person Re-Identification (UVI-ReID) methods: OTAL [49], CCLNet [50], PGM [51], GUR [52], SDCL [11], and MMM [12], which are trained exclusively on the target domain. The second group consists of three Source-Free Domain Adaptation (SFDA) methods: CMT [53], LEAD [15], and DRU [54], which involve training a source model and performing domain adaptation following the authors' recommended procedures. Based on the results shown in Tables 1 and 2, several important observations can be made:

- SVIP consistently achieves superior performance across all benchmarks under both the *Basic Setting* and *Weather Setting*, highlighting its strong generalization capability. In some scenarios, the R-1 accuracy exceeds that of the second-best method by over 7%, reflecting its robustness in handling challenging conditions.
- Effective exploitation of source domain knowledge plays a crucial role in enhancing the understanding of the target domain, which largely contributes to SVIP's superiority over existing UVI-ReID methods that rely solely on target domain data.
- SFDA baseline methods are generally constrained by the assumption of shared classes between the source and target domains, which hinders their ability to address the challenge of non-overlapping classes (identities) in VI-ReID. This limitation often results in suboptimal performance, and in certain cases, even inferior outcomes compared to UVI-ReID.

4.3 Ablation Study

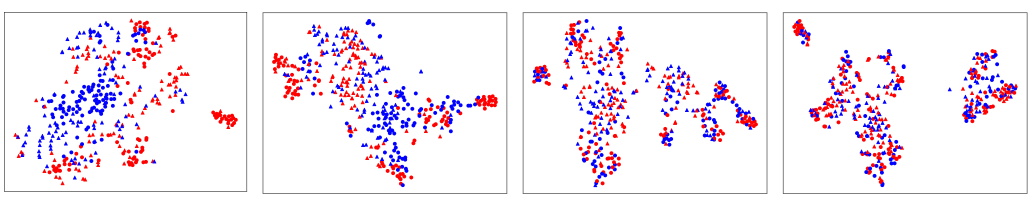
We conduct comprehensive ablation studies to assess the individual contributions of the key components in SVIP, namely SGCL, RTL, and SCCA, to the overall adaptation performance. Specifically, we progressively remove each component and evaluate the resulting performance degradation on the target domain. As reported in Table 3, the exclusion of any single component leads to a notable drop in re-identification accuracy, which clearly demonstrates the importance of each mechanism in enhancing domain adaptation.

Table 3: Ablation studies on *Basic Setting*, where RegDB dataset as target domain.

SGCL	RTL	SCCA	SYSU-MM01 → RegDB				LLCM → RegDB			
			V2T		T2V		V2T		T2V	
			R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
✓			85.26	75.64	82.13	72.45	84.32	76.68	81.05	71.32
	✓		83.71	73.82	80.59	70.28	82.47	74.13	79.24	69.85
		✓	81.95	71.36	78.43	68.17	80.63	71.89	77.52	67.41
✓	✓		87.39	77.25	84.16	74.83	86.21	78.04	83.17	73.62
	✓	✓	88.17	78.06	85.04	75.91	87.35	78.97	84.26	76.13
✓		✓	86.82	76.43	83.27	73.85	85.74	77.32	82.39	72.68
✓	✓	✓	**90.13**	**81.26**	**86.92**	**77.89**	**90.20**	**81.64**	**86.54**	**78.51**

4.4 Visualization on the Domain Adaptation

To systematically evaluate our method's capability in mitigating domain shift, we conduct t-SNE visualization under the *Basic Setting* with RegDB as target domain. As shown in Figure 3, we analyze feature distributions across four configurations using five identities (20 images each) from source and target domains. Based on the experimental results, we can make the following observations: 1) Source Only exhibits severe domain gap with separated source (red) and target (blue) clusters; 2) SGCL partially reduces domain discrepancy through confidence-weighted dual clustering but retains scattered distributions due to insufficient cross-modal alignment; 3) SGCL+RTL enhances structural stability by aligning feature residuals, though noise persists from imperfect sample filtering; 4) The full framework (SGCL+RTL+SCCA) achieves tight cross-domain clustering through SCCA-guided structural consistency constraints and confidence-aware filtering. This is clearly shown in the t-SNE visualization, where corresponding blue circles and triangles are closely clustered, reflecting accurate alignment between matched samples. The progressive improvements across different stages further validate the effectiveness of our method in establishing stable cross-modal correspondences and effectively mitigating domain shift.



(a) Source Only. (b) SGCL. (c) SGCL+RTL. (d) SGCL+RTL+SCCA.

Figure 3: The **Red** and **Blue** denote source and target domains, respectively, while ● and ▲ denote visible and infrared images (best viewed in color).

5 Conclusion

This paper introduces a novel framework for source-free domain adaptation in visible-infrared person re-identification (VI-ReID), termed SVIP. Without access to source data, SVIP effectively leverages the pretrained source model through three tailored components: 1) a Source Guided Contrastive Learning (SGCL) mechanism that improves supervision quality via dual-model clustering and confidence-weighted supervision; 2) a Residual Transfer Learning (RTL) mechanism that distills transferable knowledge through intermediate feature alignment; and 3) a Structural Consistency Guided Cross-modal Alignment (SCCA) mechanism that exploits semantic structure from the source model to guide cross-modal association in the unpaired target domain. These designs collectively enhance target model adaptation under the SFDA setting. Future extensions include generalizing SVIP to other multi-modal retrieval tasks (e.g., video-text ReID), and exploring continual adaptation in streaming settings with evolving target domains.

https://doi.org/10.52202/085713-0690

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62472295, 62176171, U24B20174, and 62372315; in part by the Fundamental Research Funds for the Central Universities under Grants CJ202303 and CJ202403; in part by the Sichuan Science and Technology Planning Project under Grants 24NSFTD0130, 2024ZDZX0004, and 2024NSFTD0049; in part by the Chengdu Science and Technology Project under Grant 2023-XT00-00004-GX; and in part by the TCL Science and Technology Innovation Fund.

References

[1] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021.

[2] Mouxing Yang, Zhenyu Huang, and Xi Peng. Robust object re-identification with coupled noisy labels. *International Journal of Computer Vision*, pages 1–19, 2024.

[3] Honghu Pan, Wenjie Pei, Xin Li, and Zhenyu He. Unified conditional image generation for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 2024.

[4] Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18621–18632, 2023.

[5] Zhenyu Cui, Jiahuan Zhou, and Yuxin Peng. Dma: Dual modality-aware alignment for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 2024.

[6] Zhiqi Pang, Chunyu Wang, Lingling Zhao, Yang Liu, and Gaurav Sharma. Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[7] Yongxiang Li, Yuan Sun, Yang Qin, Dezhong Peng, Xi Peng, and Peng Hu. Robust duality learning for unsupervised visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 20:1937–1948, 2025.

[8] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6112–6121, 2019.

[9] Hao Feng, Minghao Chen, Jinming Hu, Dong Shen, Haifeng Liu, and Deng Cai. Complementary pseudo labels for unsupervised domain adaptation on person re-identification. *IEEE Transactions on Image Processing*, 30:2898–2907, 2021.

[10] Yan Jiang, Xu Cheng, Hao Yu, Xingyu Liu, Haoyu Chen, and Guoying Zhao. Domain shifting: A generalized solution for heterogeneous cross-modality person re-identification. In *European Conference on Computer Vision*, pages 289–306. Springer, 2024.

[11] Bin Yang, Jun Chen, and Mang Ye. Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16870–16879, 2024.

[12] Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-memory matching for unsupervised visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 456–474. Springer, 2025.

[13] Xiaoshuai Hao and Wanqian Zhang. Uncertainty-aware alignment network for cross-domain video-text retrieval. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5380–5389, 2017.

[15] Sanqing Qu, Tianpei Zou, Lianghua He, Florian Röhrbein, Alois Knoll, Guang Chen, and Changjun Jiang. Lead: Learning decomposition for source-free universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23334–23343, 2024.

[16] Yu Mitsuzumi, Akisato Kimura, and Hisashi Kashima. Understanding and improving source-free domain adaptation from a theoretical perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28515–28524, 2024.

[17] Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7640–7650, 2023.

[18] Shiqi Yang, Shangling Jui, Joost van de Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 35:5802–5815, 2022.

[19] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, pages 1–34, 2024.

[20] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9010–9019, 2021.

[21] Yixin Zhang, Zilei Wang, and Weinan He. Class relationship embedded learning for source-free unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7619–7629, 2023.

[22] Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8978–8987, 2021.

[23] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

[24] Shiqi Yang, Joost Van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in neural information processing systems*, 34:29393–29405, 2021.

[25] Sanqing Qu, Guang Chen, Jing Zhang, Zhijun Li, Wei He, and Dacheng Tao. Bmd: A general class-balanced multicentric dynamic prototype strategy for source-free domain adaptation. In *European conference on computer vision*, pages 165–182. Springer, 2022.

[26] Nazmul Karim, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-pang Chiu, Supun Samarasekera, and Nazanin Rahnavard. C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24120–24131, 2023.

[27] Chao Su, Huiming Zheng, Dezhong Peng, and Xu Wang. Dica: Disambiguated contrastive alignment for cross-modal retrieval with partial labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20610–20618, 2025.

[28] Jichang Li, Guanbin Li, Yemin Shi, and Yizhou Yu. Cross-domain adaptive clustering for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2505–2514, 2021.

[29] Yan Wang, Junbo Yin, Wei Li, Pascal Frossard, Ruigang Yang, and Jianbing Shen. Ssda3d: Semi-supervised domain adaptation for 3d object detection from point cloud. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2707–2715, 2023.

 https://doi.org/10.52202/085713-0690

[30] Tao Chen, Yanrong Guo, Shijie Hao, and Richang Hong. Semi-supervised domain adaptation for major depressive disorder detection. *IEEE Transactions on Multimedia*, 26:3567–3579, 2023.

[31] Jian Han, Ya-Li Li, and Shengjin Wang. Delving into probabilistic uncertainty for unsupervised domain adaptive person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 790–798, 2022.

[32] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2723–2738, 2020.

[33] Hamza Rami, Matthieu Ospici, and Stéphane Lathuilière. Online unsupervised domain adaptation for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3830–3839, 2022.

[34] Lan Yan, Wenbo Zheng, and Kenli Li. Source-free domain adaptive person search. *Pattern Recognition*, 161:111317, 2025.

[35] Xiaofeng Qu, Huaxiang Zhang, Lei Zhu, Liqiang Nie, and Li Liu. Aamt: Adversarial attack-driven mutual teaching for source-free domain-adaptive person reidentification. *IEEE Transactions on Multimedia*, 2024.

[36] Xiaofeng Qu, Li Liu, Lei Zhu, Liqiang Nie, and Huaxiang Zhang. Source-free style-diversity adversarial domain adaptation with privacy-preservation for person re-identification. *Knowledge-Based Systems*, 283:111150, 2024.

[37] Yongxiang Li, Yang Qin, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Romo: Robust unsupervised multimodal learning with noisy pseudo labels. *IEEE Transactions on Image Processing*, 2024.

[38] Yanglin Feng, Yang Qin, Dezhong Peng, Hongyuan Zhu, Xi Peng, and Peng Hu. Pointcloud-text matching: Benchmark dataset and baseline. *IEEE Transactions on Multimedia*, 2025.

[39] Yongxiang Li, Dezhong Peng, Haixiao Huang, Yizhi Liu, Huiming Zheng, and Zheng Liu. Multi-granularity confidence learning for unsupervised text-to-image person re-identification with incomplete modality. *Knowledge-Based Systems*, 315:113304, 2025.

[40] Yanglin Feng, Hongyuan Zhu, Dezhong Peng, Xi Peng, and Peng Hu. Rono: robust discriminative learning with noisy labels for 2d-3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11610–11619, 2023.

[41] Ruitao Pu, Yuan Sun, Yang Qin, Zhenwen Ren, Xiaomin Song, Huiming Zheng, and Dezhong Peng. Robust self-paced hashing for cross-modal retrieval with noisy labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19969–19977, 2025.

[42] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[43] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

[44] Yukang Zhang and Hanzi Wang. 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*, pages 2153–2162, 2023.

[45] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14308–14317, 2022.

[46] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4330–4339, 2021.

[47] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2021.

[48] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13657–13665, 2020.

[49] Jiangming Wang, Zhizhong Zhang, Mingang Chen, Yi Zhang, Cong Wang, Bin Sheng, Yanyun Qu, and Yuan Xie. Optimal transport for label-efficient visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 93–109. Springer, 2022.

[50] Zhong Chen, Zhizhong Zhang, Xin Tan, Yanyun Qu, and Yuan Xie. Unveiling the power of clip in unsupervised visible-infrared person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3667–3675, 2023.

[51] Zesen Wu and Mang Ye. 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*, pages 9548–9558, 2023.

[52] Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11069–11079, 2023.

[53] Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23839–23848, 2023.

[54] Trinh Le Ba Khanh, Huy-Hung Nguyen, Long Hoang Pham, Duong Nguyen-Ngoc Tran, and Jae Wook Jeon. Dynamic retraining-updating mean teacher for source-free object detection. In *European Conference on Computer Vision*, pages 328–344. Springer, 2024.

 https://doi.org/10.52202/085713-0690

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the Supplementary Material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The Analysis and proof are provided in the supplementary material.

https://doi.org/10.52202/085713-0690

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation and training details are clearly described for reproduction in our main paper and supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

https://doi.org/10.52202/085713-0690

Justification: The code will be released publicly after in-peer review.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiment settings are clearly presented in the paper and supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because it would be too computationally expensive for experiments involving LLMs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources are reported in the experiment settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts are discussed with the limitations.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

https://doi.org/10.52202/085713-0690

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: All datasets and models used in this paper are publicly available.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Proper citations are provided throughout the document and the licenses will be included with the code when it is released.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The document will release.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

https://doi.org/10.52202/085713-0690

Justification: We have fully disclosed the details of the use of the adopted LLMs in our supplementary material.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-0690