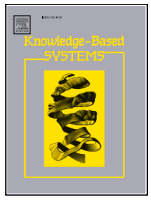
Knowledge-Based Systems 334 (2026) 115015

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Multi-level alignment network for unsupervised domain adaptive multi-modality object re-identification

Yusong Sheng [a], Yuhe Ding [b], Aihua Zheng [a,c,d], Ziqi Liu [e], Zi Wang [f,g,*], Jin Tang [b]

[a] Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui University, Hefei, 230601, China
[b] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China
[c] Anhui Provincial Key Laboratory of Intelligent Detection and Diagnosis for Traffic Infrastructure, Anhui University, Hefei, 230601, China
[d] School of Artificial Intelligence, University, Hefei, 230601, China
[e] the Kingsway College, Oshawa, L1K2H4, Canada
[f] School of Biomedical Engineering, Anhui Medical University, Hefei, 230032, China
[g] Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education, Anhui University, Hefei, 230601, China

ARTICLE INFO

Keywords:
Multi-modality object re-identification
Domain adaptation
Prototype
Distribution alignment

ABSTRACT

Existing multi-modality object re-identification methods demonstrate robust performance in complex environments, but this is predominantly contingent upon the test and training data sharing an identical distribution. However, performance degrades significantly when applied to the real world (target domains) with distribution differences from the training data (source domain). Single-modality domain adaptation methods that do not account for multi-modality domain gaps and complementary modality information do not achieve satisfactory performance. To alleviate this, we first introduce the task of unsupervised domain adaptation multi-modality object ReID, aiming to address the challenge of distribution shift in multi-modality scenarios and its impact on model performance. We further propose a novel Multi-level Alignment Network (MAN), which performs alignment strategies at the pseudo-label level, domain level, and modality level by leveraging multi-modality information consistency, multi-modality distribution discrepancy, and multi-modality information diversity. Specifically, Consistency-driven Pseudo-label Alignment (CPA) aims to mitigate the effects of pseudo-label noise from clustering by aligning pseudo-labels and filtering reliable samples based on consistency scores. Prototype-guided Domain Distribution Alignment (PDA) narrows the domain gap between the source and target domains by minimizing the distribution distance between the prototype of one domain and the instances of another domain. Margin-preserved modality distribution alignment (MMA) aligns modality distributions within the same domain by keeping the distribution of instances to a marginal distance from the prototype distribution and preserves modality diversity and complementary information. Experiments conducted on vehicle and person datasets WMVeID863, RGBNT100, RGBNT201, and Market1501-MM validate the effectiveness of the proposed method.

1. Introduction

Object re-identification (ReID) relying solely on visible modality faces significant challenges in complex visual environments [2,8]. Recently proposed multi-modality methods enhance the robustness of the feature for practical applications by incorporating various data sources (e.g., near-infrared and thermal infrared) [9,11]. Although these supervised multi-modality ReID methods perform well in various complex environments with consistent distribution, their cross-domain transfer performance remains concerning when there is a distribution mismatch between the test and training data.

We focus for the first time on the task of unsupervised domain adaptation multi-modality object ReID (UDA MMReID) as shown in Fig. 1(a). The key challenges lie in mitigating domain gaps across different modalities and effectively leveraging complementary multi-modality information under the condition that the target domain data without labels. Some UDA single-modality ReID methods, which employ joint learning and improve pseudo-labels through clustering and fine-tuning [12,18] can also be applied to the UDA MMReID task after performing multi-modality variants. However, those single-modality methods typically only need to consider domain differences between a single modality, generally by clustering to obtain pseudo-labels and using these

* Corresponding author.
E-mail addresses: sys115454324@163.com (Y. Sheng), madao3c@foxmail.com (Y. Ding), ahzheng214@foxmail.com (A. Zheng), zliu33837@gmail.com (Z. Liu), ziwang@ahmu.edu.cn (Z. Wang), tangjin@ahu.edu.cn (J. Tang).

https://doi.org/10.1016/j.knosys.2025.115015
Received 4 August 2025; Received in revised form 11 November 2025; Accepted 27 November 2025
Available online 9 December 2025

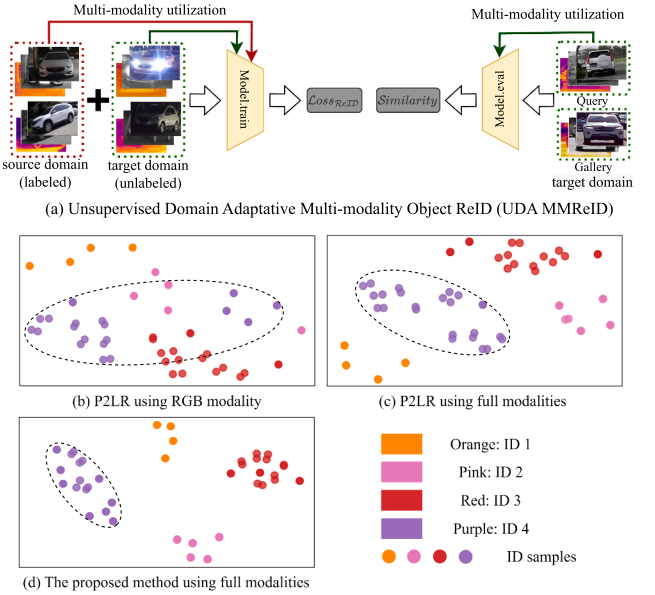


Fig. 1. (a) The processing of the newly proposed UDA MMReID task. (b-c) The visualization of feature distribution extracted by the P2LR [1] on single-modality and multi-modality in the "MSVR310 → WMVeID863" transfer task. The feature distribution of ID 4 in the target domain is decentralized. (d) The visualization of feature distribution extracted by the proposed MAN on multi-modality data. The feature distribution of ID 4 in the target domain is more concentrated.

pseudo-labels as supervisory signals to train the model. As shown in Fig. 1(b), the single-modality cross-domain re-identification method P2LR [1] can also be adapted for application in the multi-modality cross-domain scenario. Although there is some performance improvement, the overall results are not satisfactory as shown in Fig. 1(c). These methods face some major limitations when focused on UDA MMReID: they fail to fully exploit the complementary information of multi-modality and the domain discrepancies between different modalities.

To alleviate the above limitations, we propose the Multi-level Alignment Network (MAN), which reduces domain distribution differences and multi-modality discrepancies while focusing on utilizing multi-modality complementary information to enhance the model's performance in the target domain. Our MAN framework implements alignment strategies at the pseudo-label level, domain level, and modality level by leveraging multi-modality information consistency, multi-modality distribution discrepancy, and multi-modality information diversity respectively to enhance the performance of the source model on the target domain.

Specifically, consistency-driven pseudo-label alignment (CPA) aims to reduce label noise by exploiting multi-modality information consistency. CPA first aligns the clustered labels obtained from different modalities' features. Then, these aligned labels are filtered based on consistency scores and reliable samples with higher consistency are selected. This is used to explore the consistency among multi-modality information. Prototype-guided domain distribution alignment (PDA) aims to mitigate the domain distribution bias between modalities by exploiting the multi-modality distribution variability, and PDA first performs prototype-instance distribution alignment both intra-modality and inter-modality. The goal is to align the multi-modality distributions of the two domains into a unified feature space. Margin-preserving modality distribution alignment (MMA) aims to align different modalities within the same domain to a marginal distance by exploiting the multi-modality information variability while appropriately preserving the modality diversity. MMA aligns the centers of the modality distributions using an unsupervised multi-modality alignment loss (UMA loss) designed with a margin distance. Using the above three alignment strategies, the distribution of features on the target domain is more concentrated and has higher ID discrimination as shown in Fig. 1(d). This indicates that our MAN can better utilize multi-modality information that cannot be considered by single-modality methods and effectively reduce domain gaps, leading to improved performance.

In summary, our contributions are as follows:

- We first introduce an unsupervised domain adaptation multi-modality object re-identification (UDA MMReID) task and a novel Multi-level Alignment Network (MAN) to alleviate the performance degradation in the multi-modality target domain due to distribution shift.
- We introduce Consistency-driven Pseudo-label Alignment (CPA), aiming to mitigate the pseudo-label noise problem from clustering on unsupervised target domains using multi-modality information.
- We introduce Prototype-guided Domain distribution Alignment (PDA) and Margin-preserving modality distribution Alignment (MMA), aiming to utilize modality complementary information and alleviate domain differences by interleaving prototype instance distribution alignment.
- Experiments on five benchmarks for UDA MMReID. The experimental results compared with single-modality approaches fully validate the effectiveness of the proposed method.

2. Related work

2.1. Multi-modality object ReID

For multi-modality person ReID, Zheng et al. propose PFNet [11], which pioneered a progressive fusion network that learns to move from single-modality to multi-modality, and from local to global views, mentioning that robust multi-modality features can be learned even in the absence of certain modalities. To address the issue of ignoring modality-specific information, IEEE [10] proposes three interaction modules that enable the network to learn modality-specific features for each modality. For multi-modality vehicle ReID, HAMNet [10] is the first to propose an end-to-end learning framework that automatically fuses different spectral features for robust vehicle ReID under both day and night conditions. To overcome modality and sample differences, CCNet [19] proposes a novel cross-consistency network that generates discriminative multi-spectral feature representations for vehicle ReID. HViT [20] introduces a hybrid visual transformer approach to reduce modality-induced feature bias. To address the crucial issue of multi-modality complementary information fusion, the subsequent PHT [21] effectively fuses multi-modality information at the feature level using a transformer with random mixing enhancement and feature mixing mechanisms.

Although these methods have shown good results, they primarily focus on leveraging multi-modality information under supervised conditions. In contrast, our approach focuses on transferring knowledge from the source domain in the absence of labeled data in the target domain.

2.2. UDA single-modality ReID

UDA ReID aims to address the domain gap between the source and target domains without relying on ID labels in the target domain, using data from a single modality.

Self-training methods [22] show more promise, where a model is trained on supervised data from the source domain, and then alternates between clustering and fine-tuning in the target domain [12,14,23,28]. The introduction of this procedure in ReID, particularly by PUL [22], accelerate the development of self-training methods. MMT [12] introduces the mean teacher model, refining pseudo-labels through mutual learning. Later, Zhai et al. [29] use three networks for mutual mean

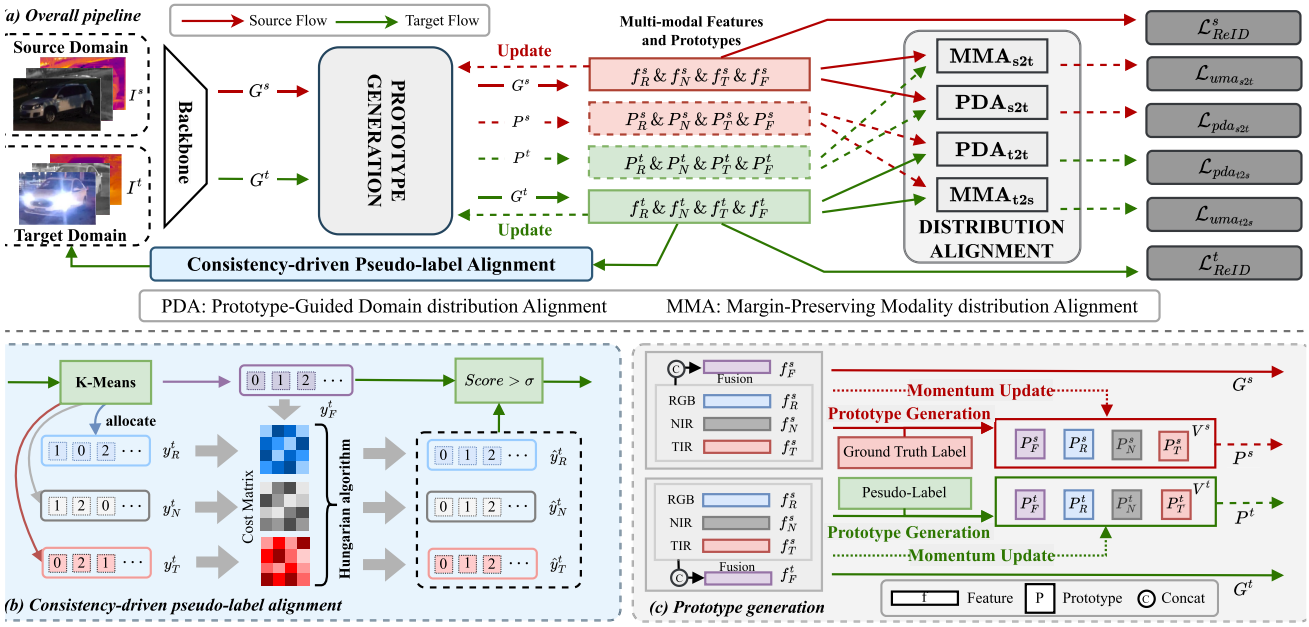


Fig. 2. An illustration of our proposed MAN. First, features from different domains and modalities, as well as fusion features, are extracted and jointly trained using a domain-shared three-branch backbone network pre-trained on the source domain. Then, the Consistency-Driven Pseudo-Label Alignment (CPA) strategy is employed to select samples with consistency scores above a threshold, assign pseudo-labels, and add them to the reliable sample set. Subsequently, Prototype-Guided Domain distribution Alignment (PDA) combined with Margin-Preserving modality distribution Alignment (MMA) is used to enhance the diversity of intra-domain modality information while mitigating modality discrepancies.

learning. Considering the impact of noisy pseudo-labels on model robustness, some methods [1,23,25,26,30,31] focus on pseudo-label evaluation and quality control. UNRN [25] evaluates the uncertainty by checking the consistency between the teacher and student model outputs. GLT [26] uses group-aware label propagation to explicitly correct noisy labels. Han et al. [1] attempt to improve pseudo-labels using progressive strategies with uncertainty. SECRET [23] refines pseudo-labels from different feature spaces to increase consistency. DKD-MPL [30] mitigated the deviation of supervision by exploiting global and local complementary knowledge between different views of pseudo-labels. **Joint learning** methods [15,16,32,33] focus on combining data from the source domain with ground-truth labels with unlabeled data from the target domain. IDM [15] proposes an intermediate domain module that generates intermediate representations linking the source and target domains for knowledge transfer. Pang et al. [33] proposed a camera invariant feature learning (CIFL) framework forced to learn camera-invariant features. To address both inter-domain transfer and intra-domain shift, Luo et al. [16] introduce camera-aware perception and cross-domain mixing to enhance neighborhood consistency, allowing a smooth transition between the source and target domains. To mitigate the error accumulation caused by pseudo-label noise, Li et al. [34] propose a dual-stream adversarial learning strategy, identity-related features are disentangled from images to obtain discriminative yet domain-invariant features. Qi et al. [35] further explored sample similarity in sample pair space based on the observation that the distribution gap is smaller in sample pair space than in sample instance space. The subsequent Chen et al. [36] utilize multi-center memory to capture different camera information for each identity and align the information of the same identity across different cameras. Wang et al. [31] focus on enhancing feature discriminability to address the contradiction between intra-class diversity and pseudo-label accuracy.

Following them, we extend the joint learning concept, aiming to resolve domain misalignment and modality-specific misalignment in multi-modality scenarios and decide to leverage multi-modality information to mitigate label noise.

3. Methodologies

3.1. Overview

In the newly proposed unsupervised domain adaptive multi-modality object re-identification task, following the single-modality setup, it is assumed that we can obtain a total of z^s labeled samples for l IDs from the source domain D^s for each modality $\{[I^s_{R,i}, I^s_{N,i}, I^s_{T,i}, y^s_i]|_{i=1}^{z^s}\}$ where $I^s_i \in I^s$, $y^s_i \in Y^s$, and R, N, T stand for RGB, NIR, TIR, respectively. We will keep this expression for later. And in the target domain D^t we can obtain k IDs totaling z^t unlabeled samples $\{[I^t_{R,i}, I^t_{N,i}, I^t_{T,i}]|_{i=1}^{z^t}\}$ where l and k are not necessarily equal. As well as the pseudo-labels $\{\hat{y}^t_{R,i}, \hat{y}^t_{N,i}, \hat{y}^t_{T,i}|_{i=1}^{z^t}\}$ of the target domain obtained by clustering of K-means. We use triple-branch ResNet50 [37] as the backbone network E^s and add a fully connected dimensions layer $l+k$ as a hybrid classifier followed by a softmax activation function. Firstly, pre-train it on the source domain data using the common ReID loss.

Fig. 2(a) illustrates the overall framework of our method. Our Multi-level Alignment Network (MAN) adopts a joint training strategy for the labeled source domain and the unlabeled target domain, where each mini-batch consists of b source samples and b target samples. The goal is to maintain the performance of the source domain while effectively utilizing the unlabeled target data to help the model better adapt to the target domain. MAN has two key components: Consistency-driven Pseudo-label Alignment (CPA) and Distribution Alignment. The Distribution Alignment consists of three parts: prototype generation as a prerequisite, followed by Prototype-guided Domain distribution Alignment (PDA) and Margin-preserving modality distribution Alignment (MMA). First, the features extracted by the backbone (i.e., G^s, G^t) are fed into CPA to obtain a reliable sample set $D^t_{reliability}$ by label alignment strategies and consistency scores. Next, the generated prototypes (P_R, P_N, P_T) and the extracted features (f_R, f_N, f_T) are passed into PDA and MMA, which restrict the consistency of the domain distribution and the consistency of the modality distribution. Finally, the resulting features (f_R, f_N, f_T) are passed through the classifier to calculate the ReID loss.

3.2. Consistency-driven pseudo-label alignment

Based on the pre-trained three-branch source domain model E^s that we have obtained, we first initialize the target domain model E^t with it. Then, we follow previous unsupervised ReID methods and assign pseudo-labels to the unlabeled target domain data samples through clustering. Since the images $\{I_{R,i}, I_{N,i}, I_{T,i}\}$ of each sample from the three modalities are paired, they share the same ID label y_i. Subsequently, we conduct a preliminary validation of the impact of different modalities on the accuracy of clustering results, as shown in Fig. 3. The different information contained in each modality leads to differences in clustering performance. The fusion modality feature f_F^t contains richer information, resulting in a higher accuracy of the label compared to the accuracy of individual modality features.

3.2.1. Fusion feature clustering

Higher-quality clustering results can be achieved due to the fusion of multiple modality features by splicing them together. Therefore, we choose to concatenate the features to obtain the fusion feature f_F^t for clustering:

$$Cluster_F = KM(Cat(f_R^t, f_N^t, f_T^t)), \tag{1}$$

where KM represents clustering of K-means algorithm, Cat denotes the concatenation operation, and f_R^t, f_N^t, f_T^t are the modality-specific features extracted by the backbone network. We then assign pseudo-labels y_F^t to each sample:

$$PL_F = \left[y_{F,1}^t, y_{F,2}^t, \ldots, y_{F,z^t-1}^t, y_{F,z^t}^t\right], \tag{2}$$

where F represents the fusion modality and k is the number of clusters obtained by clustering of the K-means algorithm. However, the pseudo-labels y_F^t obtained by fusion feature clustering alone still have poor accuracy, and directly using these labels for supervised training still affects the performance of the model. To mitigate the performance degradation caused by pseudo-label noise, we obtain the clustering results for each modality and leverage the modality-specific information to eliminate certain unreliable labels.

3.2.2. Modality feature clustering

Similar to the fusion feature clustering part, we use the K-means algorithm to cluster the features of each modality individually in order to obtain clustering results for each of the three modalities and assign pseudo-labels y_R^t, y_N^t, y_T^t accordingly For each sample I_i^t:

$$\begin{aligned} Cluster_M &= KM(f_M^t), \\ PL_M &= \left[y_{M,1}^t, y_{M,2}^t, \ldots, y_{M,z^t-1}^t, y_{M,z^t}^t\right]. \end{aligned} \tag{3}$$

Each modality $I_{M,i}^t$ of each sample I_i^t is assigned the current intra-modality ID, given that the unsupervised clustering algorithm can assign different labels to different modalities of the same sample. This will separate the three modality picture pairs of a sample, negatively affecting the separation. Therefore, we further propose pseudo-label alignment complemented by a consistent selection strategy to address this problem so that the different modality pictures of each sample remain paired after global and modality clustering.

3.2.3. Pseudo-label alignment and consistency selection

As shown in Fig. 2(b), we align each modality label PL_R, PL_N, PL_T to the fusion feature label PL_F with higher accuracy as a benchmark. Specifically, we align the individual single-modality pseudo-labels y_M^t to the fusion modality pseudo-labels y_F^t, so that the pseudo-labels assigned to the same ID samples are the same, thus realizing the preservation of modality picture pairs. Subsequently, the consistency of the modality information (which can be viewed as the same samples in all three modalities) is utilized to achieve pseudo-label denoising.

Specifically, we compute the cost matrix between labels using the label assignment PL_F of the fusion modality and the label assignment

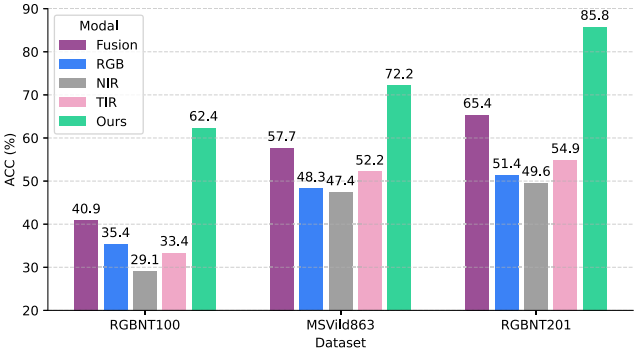


Fig. 3. The clustering accuracy of different modalities. Any single modality feature used for clustering performs worse than the fusion modalities. However, after refinement with our CPA, the accuracy is further improved.

$\{PL_M \in \mathbb{R}^k, M \in \{R, N, T\}\}$ of the individual modality. By iterating over all the cluster pairs (i, j), we calculate the number of differences (that is, the total number of mismatched samples) between the samples belonging to cluster i in PL_F and the samples belonging to cluster j in PL_M, and this difference is the matching cost. Subsequently, the Hungarian algorithm [38] is solved to find pairs of clusters with the lowest matching cost. The Hungarian algorithm is able to achieve one-to-one matching, while the K-Means-based clustering always obtains the same number of clusters, which ensures that the single-modality labels y_M^t of each sample can always find their counterparts in the fusion modality y_F^t:

$$\begin{aligned} &\hat{y}_i^t(M) = Match(y_{M,i}^t, y_{F,j}^t), i, j = 1, 2, \ldots, z^t, \\ &\hat{PL}_M = \left[\hat{y}_{M,1}^t, \hat{y}_{M,2}^t, \ldots, \hat{y}_{M,z^t-1}^t, \hat{y}_{M,z^t}^t\right]. \end{aligned} \tag{4}$$

Based on the pseudo-label alignment, we obtain a reliable sample set by means of a consistent selection strategy. Specifically, we just need to obtain the fusion pseudo-label y_F^t for each sample I_i^t and the pseudo-label $\hat{y}_M^t$ after the three modality alignments, and pass the equation:

$$score_i = \frac{\sum_{M_i \neq M_j}^{\{F,R,N,T\}} mathds1(\hat{y}_i^t(M_i) = \hat{y}_i^t(M_j))}{\sum_{M_i \neq M_j}^{\{F,R,N,T\}} mathds1,} \tag{5}$$

to obtain a label consistency score. Here $mathds1$ represents the indicator function. When the score is above the threshold σ, we consider this sample's pseudo-label to be reliable and add it to the set of reliable samples $D_{reliability}^t$:

$$D_{reliability}^t = \{\hat{I}_i^t \mid \hat{I}_i^t \in D^t \ and \ score_i > \sigma\}. \tag{6}$$

As the process of training the model using the set of reliable samples with denoised and refined pseudo-labels proceeds, the number of reliable samples gradually increases and the negative impact of incorrect labels on the model is reduced.

3.3. Prototype generation and distribution alignment

3.3.1. Prototype generation

We encourage the model to focus more on domain-invariant and modality-invariant information through interleaved prototype and instance distribution alignment. Instances reflect intra-class distinctions, while prototypes treat different classes equally, which can effectively alleviate the class imbalance problem [39]. As shown in Fig. 2(c), we begin by introducing prototypes. Traditional centroid-based methods treat all samples equally. For example, some methods [39,40] compute the centroid by taking the mean of all samples, while they neglect the fact that the confidence of the clustering labels for each sample is different. As a result, we convert the distance between each sample and its

class centroid into scores and then compute a weighted sum to form the class prototype. Compared with traditional centroids, our approach can mitigate the centroid shift caused by noisy boundary samples under unsupervised settings, thereby improving the effectiveness of domain distribution alignment.

For the source domain, we directly compute the feature centers of each category using the real labels to generate the initial source domain prototypes $P^s_{M,i}$:

$$\begin{aligned} w^s_{M,i} &= Softmax\left(1 - Cos\left(f^s_{M,i}, \frac{1}{c}\sum_{j=1}^{c} f^s_{M,i}\right)\right) \\ P^s_{M,i} &= \frac{1}{c}\sum_{j=1}^{c} w^s_{M,i} * f^s_{M,i},\ i = 1, 2, \ldots, l, \end{aligned} \tag{7}$$

here, c denotes the number of samples per class and Cos represents the cosine distance. for the target domain, we use the reliable samples $D^t_{reliability}$ refined by CPA and the pseudo-labels $\hat{PL}_M$ to generate class centroids the initial target domain prototypes $P^t_{M,i}$:

$$\begin{aligned} w^t_{M,i} &= Softmax\left(1 - Cos\left(f^t_{M,i}, \frac{1}{c}\sum_{j=1}^{c} f^t_{M,i}\right)\right) \\ P^t_{M,i} &= \frac{1}{c}\sum_{j=1}^{c} w^t_{M,i} * f^t_{M,i},\ i = 1, 2, \ldots, k. \end{aligned} \tag{8}$$

It is worth noting that the extracted target domain features may be more dispersed in distribution due to the domain gap. However, prototypes are less sensitive to outliers and can help mitigate the impact of difficult edge cases.

Based on the obtained prototypes, we maintain two prototype banks V^s and V^t, which are used to store prototypes of different modalities from both domains:

$$V^s = [P^s_1, P^s_2, \ldots, P^s_l],\ V^t = [P^t_1, P^t_2, \ldots, P^t_k], \tag{9}$$

where P_i represents the stored prototype, with each modality having one class prototype:

$$P_i = \{P_{F,i}, P_{R,i}, P_{N,i}, P_{T,i}\}. \tag{10}$$

During training, the prototypes stored in the prototype banks V^s and V^t are updated with momentum m after each batch:

$$\begin{aligned} V^s &\leftarrow \lambda V^s_i + (1-\lambda) f^s, \\ V^t &\leftarrow \lambda V^t_i + (1-\lambda) f^t, \end{aligned} \tag{11}$$

where λ is the momentum updating factor. We believe that the distribution of prototypes can effectively represent the overall domain distribution, but it completely ignores intra-class distribution information and does not accurately represent data located at the periphery of clusters. Therefore, we perform a prototype-instance-level alignment between the two domains.

The challenge of ID non-overlap between source and target domains for cross-domain ReID makes it difficult to apply previously studied unsupervised classification schemes in ReID tasks. On the one hand, there are modality differences across modalities, and we propose Prototype-guided Domain distribution Alignment (PDA) to mitigate the domain differences through intra-modality and inter-modality domain distribution alignment. On the other hand, we propose margin-preserving modality distribution alignment (MMA) to improve the utilization of modality information consistency.

3.3.2. Prototype-guided domain distribution alignment

During the joint training process, we obtain reliable samples through the previously proposed CPA. Suppose that there is a large difference between the source and the target. In that case, some domain edge samples cannot be included in the reliable sample set, which prevents the model from utilizing these remaining target domain samples. Therefore, we would like to utilize the multi-modality distribution variability to reduce the gap between the two domains so that the model transitions smoothly from the source domain to the target domain.

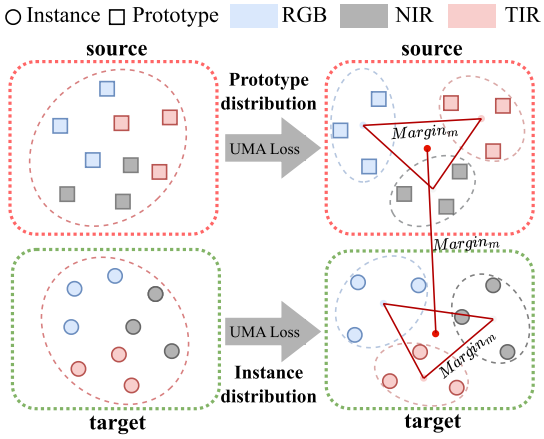


Fig. 4. Illustration of margin-preserving modality distribution alignment. We take the workflow of MMA_{t2s} as an example, where we align the source domain prototype centers and target domain instance centers within the domain. This helps to mitigate the variability of the modality distribution while retaining some modality-specific information.

Combined with the prototype bank V^s, V^t obtained previously, we decide to utilize alignment of prototype-level to instance-level across domains to reduce domain differences in the distribution and smooth the learning of domain-independent information. Specifically, we use the target domain prototype P^t_M as an anchor and separately apply maximum mean discrepancy (MMD) loss in intra-modality and inter-modality to constrain the instance features f^s_M in the source domain:

$$MMD_{I^s_M P^t_M} = K_{I^s_M I^s_M} + K_{P^t_M P^t_M} - 2K_{I^s_M P^t_M}, \tag{12}$$

where $M \in \{F, R, N, T\}$, The kernel mean $K_{I^s_M P^t_M}$ in the above formula:

$$K_{I^s_M P^t_M} = \frac{1}{lk}\sum_{i=1}^{l}\sum_{j=1}^{k} kernel(f^s_{M,i}, P^t_{M,j}), \tag{13}$$

where $kernel$ denotes the Gaussian kernel, $kernel(x, y) = \langle \phi(x), \phi(y) \rangle$, where $\phi(\cdot)$ is an implicit mapping that maps the input features to the regenerative kernel Hilbert space.

For intra-modality distribution alignment, we directly compute and optimize the MMD loss for prototype-instance between the same modality:

$$\mathcal{L}_{intra}(I^s, P^t) = \sum_{M}^{\{R,N,T\}} MMD_{I^s_M P^t_M}, \tag{14}$$

For inter-modality distribution alignment, we maintain a certain margin $margin_p$ in the domain distribution to retain some domain-specific information.

$$\mathcal{L}_{inter}(I^s_{M_i} P^t_{M_j}) = max(|margin_p - MMD_{I^s_{M_i} P^t_{M_j}}|), \tag{15}$$

where $M_i, M_j \in \{R, N, T\}$ and $M_i \neq M_j$. When using Algorithm 1 the target domain prototype P^t_M as an anchor, the process is similar to that described above. Thus, we obtain the prototype-guided domain alignment loss $\mathcal{L}_{PDA}$:

$$\begin{aligned} \mathcal{L}_{PDA} =& \mathcal{L}_{intra}(I^s, P^t) + \mathcal{L}_{intra}(P^s, I^t) \\ &+ \mathcal{L}_{inter}(I^s, P^t) + \mathcal{L}_{inter}(P^s, I^t). \end{aligned} \tag{16}$$

Algorithm 1 Multi-level domain adaptation.

Require:
Source domain data: $D^s = \{[I_R^s, I_N^s, I_T^s]\}$
Target domain data: $D^t = \{[I_R^t, I_N^t, I_T^t]\}$
Shared encoder: E^s, E^t
Ensure: Adapted feature encoder for both domains
for each epoch **do**
  Source Processing:
  Extract features: $f_R^s, f_N^s, f_T^s = E^s(I_R^s, I_N^s, I_T^s)$
  Initialize source prototype bank V^s via Eq. (7)
  Target Processing:
  Extract features: $f_R^t, f_N^t, f_T^t = E^t(I_R^t, I_N^t, I_T^t)$
  Filtering reliable pseudo-labels $\hat{PL}_M$ via Eq. (4)
  Initialize target prototype bank V^t via Eq. (8)
  for each iteration **do**
    Domain distribution alignment:
    Calculate Intra-modality loss $\mathcal{L}_{intra}$ via Eq. (14)
    Calculate Inter-modality loss $\mathcal{L}_{inter}$ via Eq. (15)
    Calculate total loss $\mathcal{L}_{PDA}$ via Eq. (16)
    modality distribution alignment:
    Calculate total loss $\mathcal{L}_{MMA}$ via Eq. (20)
    Optimizer model via Eq. (23)
    Update prototypes by momentum via Eq. (11)
  end for
end for

3.3.3. Margin-preserving modality distribution alignment

In the multi-modality case, exploiting the complementary information of different modalities is also important to facilitate effective domain migration modeling. We propose the Unsupervised Multi-modality Alignment Loss (UMA Loss), which keeps the marginal distance between the centers of different modalities at a set $margin_m$. The aim is to take full advantage of the differences in modality information and consider the greatest utilization of the information of each modality. This also improves the clustering accuracy of each modality, which is conducive to screening more reliable samples.

As shown in Fig. 4, we align the source domain prototype P_M^s with the target domain instances I_M^t for multi-modality alignment. First, we compute the center $C_{P_M^s}$ of each modality class prototype in the source domain and instance centers $C_{I_M^t}$ in the target domain :

$$C_{P_M^s} = \frac{1}{l}\sum_{i=1}^{l} P_{M,i}^s, C_{I_M^t} = \frac{1}{k}\sum_{i=1}^{k} f_{M,i}^t. \tag{17}$$

Next, the pairwise cosine distance or $L2$ distance between modality centers is computed, and the distance is kept within the margin $margin_m$:

$$\begin{aligned} &d(C_{P_{M_i}^s}, C_{P_{M_j}^s}) = \|C_{P_{M_i}^s} - C_{P_{M_j}^s}\|^2, \\ &Dist_{P^sP^s} = \max(|margin_m - d(C_{P_{M_i}^s}, C_{P_{M_j}^s})|), \end{aligned} \tag{18}$$

where $M_i, M_j \in \{R, N, T\}$ and $M_i \neq M_j$,our unsupervised multi-modality alignment loss can be expressed as:

$$\mathcal{L}_{uma}(D^t, D^s) = Dist_{I^tI^t} + Dist_{P^sP^s} + Dist_{I^tP^s}. \tag{19}$$

When aligned with the target domain prototype P_M^t and the source domain instance I_M^s, the process is similar to the above. Thus, our final margin-preserving modality distribution alignment loss is expressed as:

$$\mathcal{L}_{MMA} = \mathcal{L}_{uma}(D^t, D^s) + \mathcal{L}_{uma}(D^s, D^t). \tag{20}$$

By reducing the distance between the centers of the modality features, the data of different modalities exhibit stronger consistency in the same feature space.

With the proposed prototype-based domain distribution and modality distribution alignment strategies, the model can better learn domain-invariant and modality-invariant knowledge within the aligned distributions, while maintaining the margin distance also supports the retention of domain-specific and modality-specific information.

3.4. Objective function

As illustrated in Fig. 2(a), our objective is to optimize the feature extraction capability of the shared backbone network of the two domains. The overall loss function of the pre-trained source model can be expressed as:

$$\mathcal{L}_{ReID} = \mathcal{L}_{id} + \mathcal{L}_{triplet}, \tag{21}$$

For the source domain, we use truthful labeling supervision, and for the target domain, we use reliable pseudo-labeling $\hat{\text{PL}}_{\text{M}}$. In the supervised case, we use the cross-entropy loss with label smoothing and the ternary loss with equal weights:

$$\begin{aligned} \mathcal{L}_{id} &= \frac{1}{z}\sum_{m}^{m\in M}\sum_{i=1}^{z} y_{m,i}\log\hat{y}_{m,i}, \\ \mathcal{L}_{triplet} &= \frac{1}{z}\sum_{m}^{m\in M}\sum_{i=1}^{z} max(0, dist_m^p - dist_m^n + \alpha), \end{aligned} \tag{22}$$

where $\hat{y} = \delta_k(f(I))$, $dist_m^p = f(I_{m,i}^p) - f(I_{m,i})$, I^p represents positive samples and I^n represents negative samples. $M \in \{F, R, N, T\}$. Finally, the total loss of our framework is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{ReID} + \alpha\mathcal{L}_{PDA} + \beta\mathcal{L}_{MMA}. \tag{23}$$

4. Experiments

4.1. Dataset and evaluation protocols

4.1.1. Datasets

A total of five datasets were used in our experiments, including three vehicle datasets and two person datasets: RGBNT100 [2], MSVR310 [19], WMVeID863 [41], RGBNT201 [11], and Market1501-MM [10].

RGBNT100 [2] includes 17,250 image triples from 100 different vehicles, with 8675 image triples from 50 vehicles assigned to the training set. The remaining 50 vehicles contribute 8575 image triples to the test set gallery, from which 1715 samples are selected as the query set. **MSVR310** [19] The dataset consists of 2087 high-quality image pairs composed of three spectral modalities, derived from a total of 6261 images of 310 vehicles. The training set includes 1032 sample pairs from 155 vehicles. The gallery set contains samples from the remaining 155 vehicles, with 591 sample groups randomly selected from 52 vehicles to form the query set. **WMVeID863** [41] contains 4709 image triplets of 863 identities (IDs) of 8 camera views. The training set contains 603 randomly selected IDs, comprising a total of 3482 image triplets. For testing, there are 260 IDs with 1226 image triplets. The gallery set includes all 260 testing IDs, while the query set is randomly sampled from the gallery, consisting of 210 IDs and 959 image triplets. **RGBNT201** [11] contains 4787 tri-modality sample pairs from 201 distinct person identities. The training set consists of sample pairs from 141 different person identities, while the remaining 60 identities are allocated to the test set, serving as both query and gallery during the testing phase. **Market1501-MM** is a synthetic extension of the Market-1501 [42] dataset. The training-test split follows the original dataset: the training set contains 12,936 image pairs from 751 identities, the query set consists of 3368 image pairs from 750 identities, while the gallery set comprises 19,732 triples from the same 750 identities. Dataset-specific details can be found in the reference ([10]).

In particular, the MSVR310 [19] and WMVeID863 [41] datasets are expanded on the basis of the earlier datasets. To ensure fairness, we remove the overlapping data from these datasets. This operation makes

Table 1
Performance(%) comparison with different modalities. It is worth noting that "RGB", "NIR", "TIR" and "Full" refer to the use of individual modality data and full modality data.

Method	modality	WMVeID863 → RGBNT100		Market1501-MM → RGBNT201	
		mAP	R-1	mAP	R-1
DHCCN [43]	RGB	27.50	50.20	13.71	12.32
	NIR	17.70	35.51	11.46	9.09
	TIR	23.50	49.39	18.45	16.99
	RGB + NIR	29.02	52.01	18.23	17.46
	RGB + TIR	37.78	74.99	24.85	23.86
	NIR + TIR	34.10	70.44	21.89	21.41
	FULL	**41.09**	**69.91**	**27.75**	**26.32**
	NIR	20.29	39.48	9.60	8.13
	TIR	26.01	55.80	19.69	15.67
	RGB + NIR	34.70	63.03	15.98	13.76
	RGB + TIR	39.80	74.93	22.33	18.42
	NIR + TIR	34.04	69.56	17.69	14.35
	FULL	**45.96**	**80.47**	**21.38**	**17.22**

the remaining samples in the MSVR310 [19] dataset particularly challenging. Therefore, we only use this dataset as the source domain and not as the target domain.

4.1.2. Evaluation protocols

To evaluate the performance of our method, we conduct experiments on multi-modality person and vehicle re-identification datasets. Consistent with standard ReID tasks, we use the mean average precision (mAP) and the cumulative matching characteristic (CMC) Rank-1/5/10 (R1/5/10) to evaluate performance. Following the mainstream evaluation protocols of single-modality methods, we evaluate our method under both the UDA ReID and US ReID tasks. (1) UDA ReID Protocol: During training, we train the model on a labeled source dataset and an unlabeled target dataset. We test the model on the target dataset using mAP and CMC. (2) US ReID Protocol: To ensure a fair comparison under unsupervised settings, we pre-train the model on the source dataset and then conduct unsupervised learning on the target domain training set. This approach aligns with the thinking of some two-stage domain adaptation methods.

4.2. Implementation details

Our model is implemented using the PyTorch toolkit and experiments are performed on NVIDIA GTX3090 GPUs. For data processing, the image size of the datasets is uniformly resized to 256×128. During the model training, we use random horizontal flipping, cropping, and erasure for data augmentation [51]. The backbone network uses ResNet50 [37] pre-trained on the ImageNet classification dataset while further experimental validation is also performed on different backbones. For model optimization, we set the small batch size to 32 and use the Adam optimizer with a learning rate of 0.00035. The momentum update parameter of the prototype is kept at 0.99 to prevent too much variation. The $margin_p$ and $margin_m$ are set to 0.5 in PDA and MMA.

4.3. Analysis of multi-modality transfer effect

As shown in Table 1, we conduct experiments using two single-modality methods on vehicle and person transfer tasks to verify the impact of multi-modality information on transfer performance. Specifically, both the US method DHCCN [43] and the UDA method P2LR [1] demonstrate that as more modalities are added, the transfer performance improves in both datasets. Since different modalities contain varying amounts of information, performance varies accordingly. However, the best performance is achieved when using all modalities. In the "WMVeID863 → RGBNT100" task, the mAP improves by almost 200 % compared to the case with the lowest performing modality. This demonstrates the effectiveness of multi-modality in cross-domain tasks.

4.4. Comparison with state-of-the-art methods

We compare our approach with state-of-the-art methods on three multi-modality vehicle datasets and two multi-modality person datasets for cross-domain tasks. Notably, since we are the first to explore UDA MMReID, the comparison methods we selected are from single-modality unsupervised (US) ReID and unsupervised domain adaptation (UDA) ReID. Specifically, we implement these methods by replacing single-modality features with multi-modality fusion features. For fairness, all the following methods use ResNet50 [37] as the backbone network.

4.4.1. Comparisons on multi-modality vehicle ReID

We first compare single-modality methods with our multi-modality domain adaptation method MAN on the multi-modality vehicle re-identification task. Overall, US methods that use source domain priors generally exhibit less robustness than UDA methods. As shown in Table 2, the US methods DHCNN [43] and LP [46] achieve second-best results in large-to-small-scale transfer task "RGBNT100 → WMVeID863" and same-scale transfer task "MSVR310 → WMVeID863", respectively. Meanwhile, the UDA method P2LR [1] achieves a second-best mAP of 45.96 % in the small-to-large-scale transfer task "WMVeID863 → RGBNT100". It is worth noting that certain methods, such as LRIMV [49], rely on camera information, which cannot function effectively in datasets such as WMVeID863 [41] due to severe camera imbalance issues. Our MAN effectively utilizes multi-modality information, ensuring state-of-the-art performance. Specifically, MAN outperforms the second-best mAP results by 4.96 %, 4.9 %, and 3.15 % in the three transfer tasks of varying scales, while also providing the most competitive rank accuracy.

4.4.2. Comparisons on multi-modality person ReID

In the multi-modality person re-identification task, despite most methods being specifically designed for person ReID, their performance is underwhelming in virtual-to-real "Market1501-MM → RGBNT201" and real-to-virtual "RGBNT201 → Market1501-MM" transfer tasks. As shown in Table 3, when considering single-modality scenarios, neither the US nor the UDA methods can effectively handle multi-modality challenges.

Although DHCNN [43] and P2LR [1] achieve second-best results with mAP of 18.14 % and 27.75 %, respectively, the issue of non-robust performance across different transfer tasks remains unresolved. In contrast, our MAN leverages multi-modality information to consistently achieve the state-of-the-art performance in the multi-modality domain adaptation for person re-identification tasks. Specifically, MAN outperforms the second-best mAP results by 5.54 % and 4.84 % in the respective transfer tasks, while also delivering the best rank accuracy. We conduct ablation experiments on the mutual transfer tasks between the RGBNT100 [2] and WMVeID863 [41] to validate the effectiveness of the components. Our MAN uses ResNet50 [37] as the backbone network, under the supervision of $\mathcal{L}_{ReID}$, and improves the baseline with XBM [52], similar to IDM.

4.5. Analysis and discussions

4.5.1. Ablation study

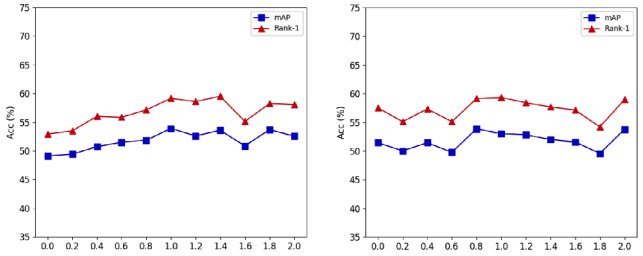
Table 4 shows the performance comparison with different components. The "Oracle" method in Table 4 uses the ground-truth label for the target (and source) domain, marking the upper bound of the UDA MMReID precision. All models, including the complete model (Ours) and those with individual components removed (-CPA, -PDA, -MMA), outperform the baseline method, demonstrating the effectiveness of the proposed approach. The complete model (Ours), integrating all components, achieves the highest results. In the setting of RGBNT100 →

Table 2
Performance comparison on multi-modality vehicle ReID benchmarks. The best and second results are in bold and underlined, respectively. "US": unsupervised methods and "UDA": unsupervised domain adaptation methods.

Methods	Reference	Type	RGBNT100 → WMVeID863				WMVeID863 → RGBNT100				MSVR310 → WMVeID863			
			mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
O2CAP [44]	TIP'22		29.78	32.42	46.7	54.21	33.79	62.04	65.25	67.76	30.04	31.32	47.80	55.68
RTMem [45]	TIP'23		36.33	39.19	52.93	60.26	35.74	70.50	72.30	73.29	40.11	43.41	56.41	63.37
LP [46]	TIP'23	US	47.87	51.83	62.82	69.41	39.16	69.21	71.49	73.18	52.17	57.14	68.68	75.46
DHCCN [43]	TCSVT'24		48.05	51.47	63.74	69.41	41.09	69.91	72.48	73.70	35.88	37.55	53.11	60.99
3C [47]	Arxiv'24		38.44	43.04	53.48	59.71	36.18	65.83	68.86	70.26	37.34	39.74	53.48	61.90
IDM [15]	ICCV'21		46.94	50.18	63.19	73.08	40.12	75.63	78.54	80.12	39.20	40.84	55.68	64.84
SECRET [23]	AAAI'22		43.89	47.62	61.90	70.88	41.74	77.55	79.94	81.63	43.89	44.14	63.55	69.96
P2LR [1]	CVPR'22		41.63	45.24	57.88	67.03	45.96	80.47	81.87	82.39	40.50	43.77	56.59	63.55
IDM + MSINet [48]	CVPR'23		41.36	44.32	55.89	64.47	37.68	75.39	77.96	79.42	39.19	40.66	56.96	64.29
LRIMV [49]	TNNLS'24		–	–	–	–	39.23	74.40	76.09	77.32	–	–	–	–
DSSM [50]	AI COMMUN'24		37.10	40.66	53.85	59.89	40.18	74.81	76.73	77.43	40.44	42.12	58.06	64.10
Ours		UDA	**53.91**	**59.16**	**71.98**	**75.46**	**50.86**	**89.15**	**89.97**	**90.44**	**55.32**	**62.09**	**72.71**	**77.47**

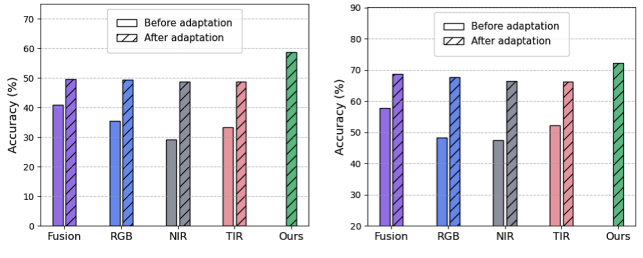
Table 3
Performance comparison on multi-modality person ReID benchmarks. The best and second results are in bold and underlined, respectively. "US": unsupervised methods and "UDA": unsupervised domain adaptation methods.

Methods	Reference	Type	RGBNT201→ Market1501-MM				Market1501-MM → RGBNT201			
			mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
O2CAP [44]	TIP'22		4.81	14.40	28.59	36.13	8.01	7.06	11.84	16.27
RTMem [45]	TIP'23		14.07	36.13	50.33	57.01	18.60	15.67	19.98	25.96
LP [46]	TIP'23	US	12.51	31.92	47.68	55.49	26.01	22.13	31.34	36.72
DHCCN [43]	TCSVT'24		8.65	24.82	41.48	49.38	27.75	26.32	36.96	44.14
3C [47]	Arxiv'24		5.79	17.10	31.83	39.64	9.32	8.73	14.23	19.74
MMT [12]	ICLR'20		12.63	32.13	46.29	52.52	15.43	12.08	21.89	27.87
IDM [15]	ICCV'21		9.82	27.35	40.65	47.33	16.24	13.64	20.81	30.74
SECRET [23]	AAAI'22		12.63	32.13	46.29	52.52	15.70	11.36	22.97	31.70
P2LR [1]	CVPR'22	UDA	18.14	39.85	55.29	61.34	21.38	17.22	28.11	34.93
IDM + MSINet [48]	CVPR'23		11.16	28.62	43.59	50.68	19.26	14.95	26.67	33.97
LRIMV [49]	TNNLS'24		16.98	39.99	58.88	67.81	–	–	–	–
DSSM [50]	AI COMMUN'24		10.67	28.44	44.03	51.40	18.07	14.00	24.64	31.22
Ours		UDA	**23.68**	**48.04**	**65.53**	**73.60**	**32.59**	**33.61**	**42.82**	**46.65**



(a) Loss weight α (b) Loss weight β

Fig. 5. Hyperparameter analysis on α and β in the "RGBNT100 → WMVeID863" task.



(a) RGBNT100 (b) WMVeID863

Fig. 6. The improvement in label accuracy for each modality before and after the MAN is applied. The highest accuracy is achieved after the CPA refinement process.

Table 4
Performance(%) comparison with different components.

Method	RGBNT100 → WMVeID863		WMVeID863 → RGBNT100	
	mAP	R-1	mAP	R-1
Oracle	62.84	68.86	69.03	94.34
Ours	**53.91**	**59.16**	**50.86**	**89.15**
- CPA	44.42	49.08	42.25	79.53
- PDA	49.64	54.76	48.93	86.59
- MMA	51.44	57.51	47.20	85.89
Baseline	40.33	44.14	39.64	76.85

Table 5
Performance(%) comparison between prototype alignment and instance alignment. Here, "Ins": Instance, "Proto": Prototype.

Method	RGBNT100 → WMVeID863				WMVeID863 → RGBNT100			
	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Ins-Ins	48.98	52.93	67.22	75.46	47.08	85.36	86.06	86.76
Proto-Proto	48.71	53.48	66.85	75.27	48.54	85.71	86.76	87.35
Ins-Proto	**53.91**	**59.16**	**71.98**	**75.46**	**50.86**	**89.15**	**89.97**	**90.44**

WMVeID863, both mAP and R-1 scores are less than 10 % lower than the upper bound (Oracle). When the CPA component is removed, the performance drops significantly in mAP (9.49 %), demonstrating the effectiveness of multi-modality pseudo-label alignment and consistency selection in mitigating pseudo-label noise. Similarly, removing the PDA component also leads to a decrease in R-1 (4.4 %), validating the effectiveness of using prototypes to mitigate intra-modality and cross-modality differ-

Table 6
Performance (%) comparison of prototype initialization methods. Here, "average": class center, "weight": weighted center.

Method	RGBNT100 → WMVeID863				WMVeID863 → RGBNT100			
	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
average	52.78	57.88	70.88	77.11	47.83	88.45	89.62	89.91
weight	**53.91**	**59.16**	**71.98**	**75.46**	**50.86**	**89.15**	**89.97**	**90.44**

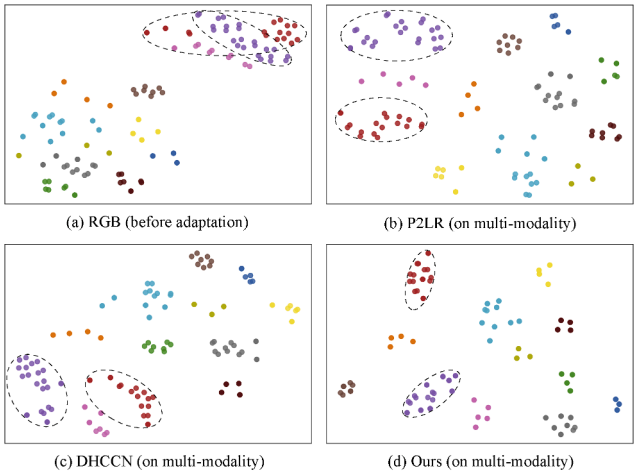


Fig. 7. The t-SNE visualization of the target domain features in the "MSVR310 → WMVeID863" task. Different colors represent different ID classes. (a) represent the feature distributions of different modalities before adaptation. (b-d) represent the feature distributions after applying existing methods and our MAN, respectively.

ences. If the MMA component is removed, there is a decline in performance as well, proving the effectiveness of margin-preserving modality distribution alignment. All these experimental results support that each module has its unique advantages, and combining them organically yields the best performance. By integrating all the components, our model achieves optimal performance. These results validate the effectiveness of our MAN in complex scenarios.

4.5.2. Effect of instance-prototype distribution alignment

As shown in Table 5, we validate the effectiveness of using prototypes in distribution alignment. The results demonstrate that both instance-instance and prototype-prototype-level distribution alignments lead to a performance improvement over the baseline with CPA. The prototype-instance level alignment, in particular, alleviates class imbalance and retains both intra-class and inter-class distribution information, resulting in further performance improvement in two different cross-domain tasks compared to the previous methods. The first and third rows in Table 5 confirm that the proposed PDA and MMA achieve better domain alignment due to the introduction of prototypes.

4.5.3. Effect of weighted prototype initialization

As shown in Table 6, we validate the effectiveness of weighted sum initialization of prototypes. The results show that the weighted summation of features for initialization is based on the scores computed from the distances between features and class centers and weights achieve better performance than the commonly used class centers.

4.5.4. Effect of hyperparameter analysis

In Fig. 5, we provide a visualization of the hyperparameters' sensitivity. Specifically, we evaluate the loss weights α and β in Eq. (23). We experiment on "RGBNT100 → WMVeID863" and evaluate the performance of mAP and Rank-1 in the target domain WMVeID863 [41].

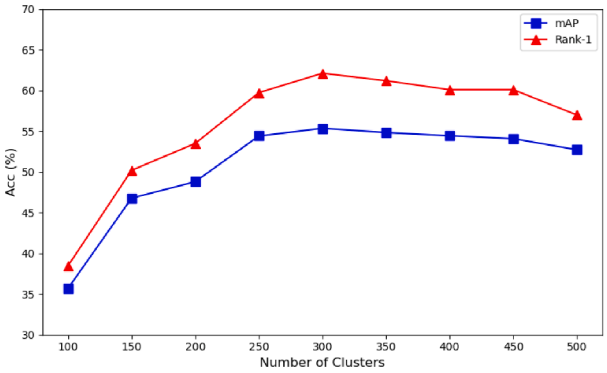


Fig. 8. Results of sensitivity analysis on the initial number of clusters in the "MSVR310 → WMVeID863" task.

As shown in Fig. 5, these hyperparameters are not very sensitive. Performance is optimal when α and β are set to 1.0 and 0.8, respectively. These hyperparameter settings are used in all experiments in this paper.

4.5.5. Effect of different backbones

To validate the effectiveness of our approach with different backbones, we conduct experiments using commonly adopted ResNet50 [37], ViT-B/16 [53], and CLIP visual encoder [54]. As shown in Table 7, our method achieves varying degrees of improvement over the Source-Only results across different backbones. Notably, since vehicles are rigid bodies whose entire visual presentation alters significantly across different angles, while ResNet [37] focuses heavily on extracting local characteristics. As a result, compared to ViT [53] and CLIP [54], ResNet proves better at identifying locally consistent clues and performs better on vehicle-related data sets. Most existing re-identification methods [55,58] have also demonstrated that models built upon ResNet [37] achieve greater performance improvements on vehicle datasets compared to those on pedestrian datasets. Simultaneously, when CLIP [54] is applied to re-identification tasks, only the visual branch is fine-tuned without corresponding text for contrastive learning, constraining its precision accuracy. Moreover, the scarce infrared data make it challenging for the sophisticated CLIP [54] to train optimally. In conclusion, we choose ResNet [37] as the baseline. In future work, we will investigate the challenges posed by limited image-text pairs and infrared data, aiming to improve the adaptability and accuracy of the ViT [53] and CLIP [54] models.

4.5.6. Complexity analysis

To evaluate the computational complexity of MAN, we measure the training time, inference time, and memory usage for each epoch on the "MSVR310 → WMVeID863", as shown in Table 8. The results demonstrate that our method has similar training and inference time compared to the joint training approach IDM [15], while consuming less GPU memory. Additionally, it outperforms mutual refinement methods like P2LR [1] and DSSM [50] in terms of speed and resource efficiency. This indicates that our approach achieves competitive performance without incurring extra computational cost during training and inference.

4.5.7. Performance with incomplete modalities

In real-world scenarios, missing modalities are common, and model robustness under such conditions is crucial. To evaluate this, we conducted experiments during the testing stage under different missing-modality scenarios. As shown in Table 9, when a single modality is absent, the accuracy of all methods exhibits varying degrees of decline. Our model still maintains relatively stable performance due to its ability to exploit complementary information across modalities. However, when the missing modality (e.g., RGB or TIR) contains more critical information, accuracy inevitably declines. Although no specific compensation

Table 7
Performance(%) comparison with different Backbone.

Method	backbone	RGBNT100 [2] → WMVeID863 [41]		WMVeID863 [41] → RGBNT100 [2]		MSVR310 [19] → WMVeID863 [41]		Market1501-MM [10] → RGBNT201 [11]		RGBNT201 [11] → Market1501-MM [10]	
		mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
Source-Only	ViT-B [53]	**32.50**	34.10	36.25	**73.06**	**34.24**	**36.45**	20.92	17.46	**14.75**	**36.02**
	CLIP [54]	25.84	23.63	**40.11**	72.30	29.51	29.85	**43.33**	**51.32**	14.22	33.79
	ResNet50 [37]	32.45	**34.43**	33.01	68.63	32.92	35.53	8.79	6.82	5.64	17.55
UDA	ViT-B [53]	50.26	54.58	41.25	81.92	52.98	58.61	25.35	21.65	25.04	49.05
	CLIP [54]	42.20	45.79	44.05	81.92	50.94	56.59	**51.59**	**52.99**	**44.15**	**71.50**
	ResNet50 [37]	**53.01**	**59.34**	**50.86**	**89.15**	**55.32**	**62.09**	32.59	33.61	23.68	48.04

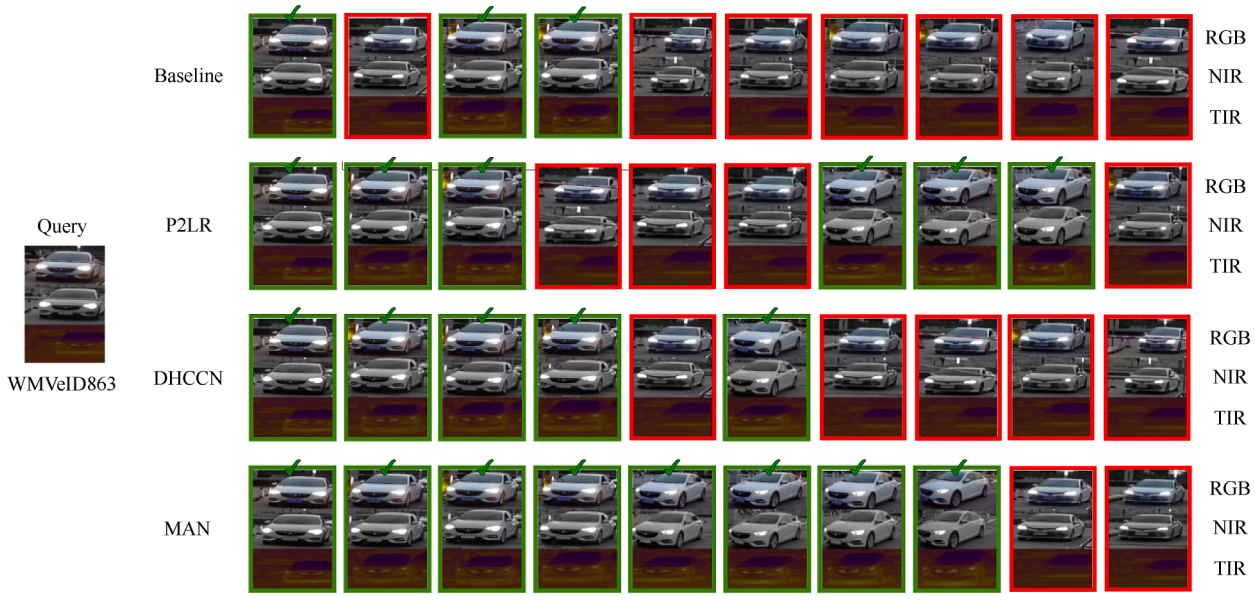


Fig. 9. Ranking results of a target domain sample query in the "MSVR310 → WMVeID863" task. The green boxes indicate the correct matchings.

Table 8
Evaluate the complexity of different methods on "MSVR310 → WMVeID863". T_{train}, T_{test} and $Memory$ represent training times, inference times and memory usage for each epoch of the method respectively.

	backbone	mAP	R-1	T_{train}	T_{test}	$Memory$
IDM [15]	ResNet50	39.20	40.84	242s	**8s**	10870M
P2LR [1]	ResNet50	40.50	43.77	253s	11s	9748M
DSSM [50]	ResNet50	40.44	42.12	220s	9s	20292M
MAN (Ours)	ResNet50	**55.32**	**62.09**	**213s**	**8s**	**8412M**

module was designed, our model still achieves competitive performance compared to other methods, owing to its overall strong adaptability to the target domain. In future work, we plan to further address this practical issue by introducing missing-modality compensation mechanisms and comparing them with existing supervised approaches.

4.5.8. Pseudo-label accuracy

As shown in Fig. 6, we show the clustering accuracy of our model for different modality features before and after training on two target datasets, including the accuracy of pseudo-labels obtained through clustering and the accuracy of reliable sample labels after CPA filtering. The results show that, before training, the quality of the features of different modalities varies, leading to inconsistent label accuracy. However, after training with our alignment strategies, the accuracy of the label for each modality improves significantly and the accuracy across modalities also becomes more consistent.

4.5.9. Feature visualization

As shown in Fig. 7, we show the visualizations of the feature distribution extracted by different methods from the source domain MSVR310 [19] to the target domain WMVeID863 [41]. Among them, (a) shows the feature distributions using a single modality. It can be observed that, regardless of the modality, the feature distributions exhibit high intra-class dispersion and low inter-class separability. In contrast, (b-d) shows the domain adaptation results using all modalities, which outperform single-modality methods. However, our MAN demonstrates better intra-class compactness and inter-class separability.

4.5.10. Number of clusters analysis

To investigate the impact of initialization and the number of clusters on model accuracy, we conducted experiments on the "MSVR310 → WMVeID863" task using multiple cluster numbers, and the corresponding results are presented in Fig. 8. As can be observed, different initial cluster numbers exert a certain influence on the final results; however, within a reasonable range, the performance variations are not particularly significant. In particular, when the cluster number is close to the optimal value that typically approximates the true number of classes in the target domain, the results remain relatively stable. Typically, we begin by dividing the total number of images by the number of images per category. We then iteratively adjust the number of images per category until the initial number of categories closely approximates the target value. This robustness can be attributed to our proposed pseudo-label filtering mechanism, which effectively mitigates the influence of noisy labels caused by hard samples. Nevertheless, when the selected cluster number deviates substantially from the optimal range, the clustering

Table 9
Performance(%) comparison with different missing modalities at test time.

		RGB		NIR		TIR		RGB + NIR		RGB + TIR		NIR + TIR	
		mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
RGBNT100 → WMVeID863	SOURCE ONLY	20.72	21.06	19.76	19.78	30.39	34.62	23.15	22.34	32.45	35.35	31.68	34.07
	BASE	24.85	26.12	22.41	22.87	34.28	38.45	27.63	27.18	37.84	40.92	36.92	38.95
	DHCCN [43]	30.67	33.59	26.89	28.74	39.24	43.87	33.75	36.89	44.38	49.63	43.12	47.28
	Ours	**38.89**	**44.14**	**33.23**	**35.71**	**45.62**	**51.28**	**42.35**	**47.44**	**53.02**	**60.26**	**51.99**	**58.79**
WMVeID863 → RGBNT100	SOURCE ONLY	22.82	45.66	16.42	35.45	21.29	46.88	24.64	51.37	32.26	67.7	27.53	60.23
	BASE	27.46	53.28	19.37	40.12	24.83	54.67	29.75	58.46	37.92	73.85	31.86	66.54
	DHCCN [43]	31.28	60.74	22.67	47.38	28.46	61.83	34.82	67.59	43.27	79.64	36.95	74.28
	Ours	**36.01**	**69.91**	**26.54**	**56.50**	**32.85**	**68.92**	**39.06**	**74.52**	**49.45**	**86.59**	**42.71**	**83.97**
MSVR310 → WMVeID863	SOURCE ONLY	22.47	21.61	19.29	19.78	29.32	33.33	25.01	24.91	33.49	36.08	32.05	34.43
	BASE	26.38	26.84	22.64	23.45	33.87	37.95	29.48	29.76	38.74	41.87	37.28	39.64
	DHCCN [43]	32.47	34.28	27.38	29.64	38.92	43.18	35.84	38.27	46.28	51.74	44.37	48.65
	Ours	**39.15**	**41.58**	**34.56**	**36.81**	**44.36**	**48.90**	**41.27**	**43.96**	**54.58**	**60.81**	**50.96**	**57.33**
Market1501-MM → RGBNT201	SOURCE ONLY	5.08	3.95	6.24	4.07	9.33	9.36	7.84	5.86	8.03	5.38	12.56	9.45
	BASE	9.27	7.84	8.92	6.95	12.68	12.87	10.38	8.64	11.28	9.47	14.38	11.28
	DHCCN [43]	13.42	11.28	9.87	7.84	13.28	13.38	14.27	12.38	15.38	13.28	16.47	14.28
	Ours	**18.82**	**18.06**	**11.08**	**9.69**	**15.78**	**14.00**	**18.64**	**17.94**	**22.31**	**19.62**	**18.66**	**17.82**

quality is severely degraded, resulting in a significant drop in accuracy. This is primarily due to the extremely low reliability of pseudo-labels when the specified number of clusters diverges significantly from the true number of classes.

4.5.11. *Ranking result visualization*

To further demonstrate the effectiveness of the proposed method, we visualize the top-10 retrieval results on the target domain WMVeID863 [41], as shown in Fig. 9, our method effectively improves the retrieval accuracy in the target domain after transferring source domain knowledge, with correctly retrieved samples ranked highly. This demonstrates that we have further leveraged multi-modality information to mitigate domain differences.

5. Conclusion

In this paper, we are the first to explore unsupervised multi-modality domain adaptation for object re-identification and propose a novel Multi-level Alignment Network (MAN). This method jointly learns from the source and target domains and further alleviates domain and modality discrepancies through Consistency-driven Pseudo-label Alignment (CPA) mitigates the pseudo-label noise obtained from clustering in the target domain, Prototype-guided Domain distribution Alignment (PDA) alleviates the differences between multi-modality domains, and Margin-preserving modality distribution Alignment (MMA) alleviates modality differences while preserving rich modality-specific information. The clustering method influences the effectiveness of subsequent alignment. In the future, we will design more suitable clustering algorithms for multi-modality data to improve the final accuracy of UDA multi-modality ReID.

CRediT authorship contribution statement

Yusong Sheng: Methodology, Investigation, Conceptualization; **Yuhe Ding:** Writing – original draft, Validation; **Aihua Zheng:** Funding acquisition, Data curation; **Ziqi Liu:** Visualization; **Zi Wang:** Writing – review & editing; **Jin Tang:** Resources.

Data availability statement

The data used to support the findings of this study are included in the paper.

Data availability

Data will be made available on request.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of competing interest

No potential conflict of interest was reported by the authors.

Acknowledgement

This research is partly supported by the National Natural Science Foundation of China (Grant No. 62372003), the University Synergy Innovation Program of Anhui Province (Grant No.GXXT-2022-036), the Natural Science Foundation of Anhui Province (Grant No.Natural Science Foundation of Anhui Province2308085Y40) and the Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education, Anhui University (Grant No. 2024A004).

References

[1] J. Han, Y.-L. Li, S. Wang, Delving into probabilistic uncertainty for unsupervised domain adaptive person re-identification, in: AAAI, 36, 2022, pp. 790–798.
[2] H. Li, C. Li, X. Zhu, A. Zheng, B. Luo, Multi-spectral vehicle re-identification: a challenge, in: AAAI, 34, 2020, pp. 11345–11353.
[3] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, S. Tian, Feature refinement and filter network for person re-identification, TCSVT 31 (9) (2021) 3391–3402. https://doi.org/10.1109/TCSVT.2020.3043026
[4] M. Zhang, M. Xin, C. Gao, X. Wang, S. Zhang, Attention-aware scoring learning for person re-identification, KBS 203 (2020) 106154.
[5] Z. Zhang, C. Lan, W. Zeng, X. Jin, Z. Chen, Relation-aware global attention for person re-identification, in: CVPR, 2020, pp. 3186–3195.
[6] P. Yan, X. Liu, P. Zhang, H. Lu, Learning convolutional multi-level transformers for image-based person re-identification, Visual Intell. 1 (1) (2023) 24.
[7] Z. Zheng, T. Ruan, Y. Wei, Y. Yang, T. Mei, VehicleNet: learning robust visual representation for vehicle re-identification, TIP 23 (2021) 2683–2693. https://doi.org/10.1109/TMM.2020.3014488
[8] X. Wang, M. Liu, F. Wang, J. Dai, A.-A. Liu, Y. Wang, Relation-preserving feature embedding for unsupervised person re-identification, TIP 26 (2024) 714–723. https://doi.org/10.1109/TMM.2023.3270636
[9] Y. Wang, X. Liu, P. Zhang, H. Lu, Z. Tu, H. Lu, TOP-ReID: multi-spectral object re-identification with token permutation, in: AAAI, 38, 2024, pp. 5758–5766.
[10] Z. Wang, C. Li, A. Zheng, R. He, J. Tang, Interact, embed, and enlarge: boosting modality-specific representations for multi-modal person re-identification, in: AAAI, 36, 2022, pp. 2633–2641.
[11] A. Zheng, Z. Wang, Z. Chen, C. Li, J. Tang, Robust multi-modality person re-identification, in: AAAI, 35, 2021, pp. 3529–3537.
[12] Y. Ge, D. Chen, H. Li, Mutual mean-teaching: pseudo label refinery for unsupervised domain adaptation on person re-identification, ICLR (2020).
[13] F. Liu, M. Ye, B. Du, Learning a generalizable re-identification model from unlabelled data with domain-agnostic expert, Visual Intell. 2 (1) (2024) 28.

[14] T. Isobe, D. Li, L. Tian, W. Chen, Y. Shan, S. Wang, Towards discriminative representation learning for unsupervised person re-identification, in: ICCV, 2021, pp. 8526–8536.
[15] Y. Dai, J. Liu, Y. Sun, Z. Tong, C. Zhang, L.-Y. Duan, IDM: an intermediate domain module for domain adaptive person RE-ID, in: ICCV, 2021, pp. 11864–11874.
[16] C. Luo, C. Song, Z. Zhang, Learning to adapt across dual discrepancy for cross-domain person re-identification, TPAMI 45 (2) (2022) 1963–1980.
[17] H. Li, N. Dong, Z. Yu, D. Tao, G. Qi, Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification, TCSVT 32 (5) (2022) 2814–2830. https://doi.org/10.1109/TCSVT.2021.3099943
[18] T. Si, F. He, Z. Zhang, Y. Duan, Hybrid contrastive learning for unsupervised person re-identification, TIP 25 (2023) 4323–4334. https://doi.org/10.1109/TMM.2022.3174414
[19] A. Zheng, X. Zhu, Z. Ma, C. Li, J. Tang, J. Ma, Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark, INFFUS 100 (2023) 101901.
[20] W. Pan, H. Wu, J. Zhu, H. Zeng, X. Zhu, H-Vit: hybrid vision transformer for multi-modal vehicle re-identification, in: IJCAI, 2022, pp. 255–267.
[21] W. Pan, L. Huang, J. Liang, L. Hong, J. Zhu, Progressively hybrid transformer for multi-modal vehicle re-identification, Sensors 23 (9) (2023) 4206.
[22] H. Fan, L. Zheng, C. Yan, Y. Yang, Unsupervised person re-identification: clustering and fine-tuning, ACM TOMM 14 (4) (2018) 1–18.
[23] T. He, L. Shen, Y. Guo, G. Ding, Z. Guo, Secret: self-consistent pseudo label refinement for unsupervised domain adaptive person re-identification, in: AAAI, 36, 2022, pp. 879–887.
[24] Q. Tian, J. Sun, Cluster-based dual-branch contrastive learning for unsupervised domain adaptation person re-identification, KBS 280 (2023) 111026.
[25] K. Zheng, C. Lan, W. Zeng, Z. Zhang, Z.-J. Zha, Exploiting sample uncertainty for domain adaptive person re-identification, in: AAAI, 35, 2021, pp. 3538–3546.
[26] K. Zheng, W. Liu, L. He, T. Mei, J. Luo, Z.-J. Zha, Group-aware label transfer for domain adaptive person re-identification, in: CVPR, 2021, pp. 5310–5319.
[27] X. Gao, Z. Chen, J. Wei, R. Wang, Z. Zhao, Deep mutual distillation for unsupervised domain adaptation person re-Identification, TIP 27 (2025) 1059–1071. https://doi.org/10.1109/TMM.2024.3459637
[28] Y. Tao, J. Zhang, J. Hong, Y. Zhu, DREAMT: diversity enlarged mutual teaching for unsupervised domain adaptive person re-Identification, TIP 25 (2023) 4586–4597. https://doi.org/10.1109/TMM.2022.3178599
[29] Y. Zhai, Q. Ye, S. Lu, M. Jia, R. Ji, Y. Tian, Multiple expert brainstorming for domain adaptive person re-identification, in: ECCV, 2020, pp. 594–611.
[30] W. Zhu, B. Peng, W.Q. Yan, Dual knowledge distillation on multiview pseudo labels for unsupervised person re-identification, TIP 26 (2024) 7359–7371. https://doi.org/10.1109/TMM.2024.3366395
[31] L. Wang, J. Huang, L. Huang, F. Wang, C. Gao, J. Li, F. Xiao, D. Luo, Attention-disentangled re-ID network for unsupervised domain adaptive person re-identification, KBS 304 (2024) 112583.
[32] S. Zhu, T. Luo, Domain-adaptive person re-identification via domain alignment and mutual pseudo-label refinement, Multimedia Syst. 30 (2) (2024) 110.
[33] Z. Pang, L. Zhao, Q. Liu, C. Wang, Camera invariant feature learning for unsupervised person re-identification, TIP 25 (2023) 6171–6182. https://doi.org/10.1109/TMM.2022.3206662
[34] H. Li, K. Xu, J. Li, Z. Yu, Dual-stream reciprocal disentanglement learning for domain adaptation person re-identification, KBS 251 (2022) 109315.
[35] L. Qi, Z. Liu, Y. Shi, X. Geng, Generalizable metric network for cross-domain person re-identification, TCSVT 34 (10) (2024) 9039–9052. https://doi.org/10.1109/TCSVT.2024.3395411
[36] K. Chen, T. Gong, L. Zhang, Camera-aware recurrent learning and earth mover's test-time adaption for generalizable person re-identification, TCSVT 34 (1) (2024) 357–370. https://doi.org/10.1109/TCSVT.2023.3285046
[37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.
[38] H.W. Kuhn, The Hungarian method for the assignment problem, Nav. Res. Logist. Q. 2 (1–2) (1955) 83–97.
[39] X. Yue, Z. Zheng, S. Zhang, Y. Gao, T. Darrell, K. Keutzer, A.S. Vincentelli, Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation, in: CVPR, 2021, pp. 13834–13844.
[40] X. Song, J. Liu, Z. Jin, Dual prototype contrastive learning with fourier generalization for domain adaptive person re-identification, KBS 256 (2022) 109851. https://doi.org/10.1016/j.knosys.2022.109851
[41] A. Zheng, Z. Ma, Z. Wang, C. Li, Flare-aware cross-modal enhancement network for multi-spectral vehicle re-identification, INFFUS 116 (2025) 102800. https://doi.org/10.1016/j.inffus.2024.102800
[42] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: ICCV, 2015, pp. 1116–1124. https://doi.org/10.1109/ICCV.2015.133
[43] Y. Li, W. Tang, S. Wang, S. Qian, C. Xu, Distribution-guided hierarchical calibration contrastive network for unsupervised person re-identification, TCSVT 34 (8) (2024) 7149–7164.
[44] M. Wang, J. Li, B. Lai, X. Gong, X.-S. Hua, Offline-online associated camera-aware proxies for unsupervised person re-identification, TIP 31 (2022) 6548–6561.
[45] J. Yin, X. Zhang, Z. Ma, J. Guo, Y. Liu, A real-time memory updating strategy for unsupervised person re-identification, TIP 32 (2023) 2309–2321.
[46] L. Lan, X. Teng, J. Zhang, X. Zhang, D. Tao, Learning to purification for unsupervised person re-identification, TIP 32 (2023) 3338–3353.
[47] M. Zheng, Y. Qu, C. Shang, L. Yang, Q. Shen, 3C: confidence-guided clustering and contrastive learning for unsupervised person re-identification, arXiv:2408.09464 (2024).
[48] J. Gu, K. Wang, H. Luo, C. Chen, W. Jiang, Y. Fang, S. Zhang, Y. You, J. Zhao, MSINet: twins contrastive search of multi-scale interaction for object reid, in: CVPR, 2023, pp. 19243–19253.
[49] S. Li, F. Li, J. Li, H. Li, B. Zhang, D. Tao, X. Gao, Logical relation inference and multiview information interaction for domain adaptation person re-identification, TNNLS 35 (10) (2023) 14770–14782.
[50] C. Tang, D. Xue, D. Chen, Doubly stochastic subdomain mining with sample reweighting for unsupervised domain adaptive person re-identification, AI Commun. (1) (37) (2024) 23-35.
[51] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: AAAI, 34, 2020, pp. 13001–13008.
[52] X. Wang, H. Zhang, W. Huang, M.R. Scott, Cross-batch memory for embedding learning, in: CVPR, 2020, pp. 6388–6397.
[53] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: transformers for image recognition at scale, arXiv:2010.11929 (2020).
[54] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: ICML, 2021, pp. 8748–8763.
[55] Y. Wang, Y. Lv, P. Zhang, H. Lu, IDEA: inverted text with cooperative deformable aggregation for multi-modal object re-identification, in: CVPR, 2025, pp. 29701–29710. https://doi.org/10.1109/CVPR52734.2025.02765
[56] Y. Wang, Y. Liu, A. Zheng, P. Zhang, Decoupled feature-based mixture of experts for multi-modal object re-identification, in: AAAI, 39, 2025, pp. 8141–8149.
[57] P. Zhang, Y. Wang, Y. Liu, Z. Tu, H. Lu, Magic tokens: select diverse tokens for multi-modal object re-identification, in: CVPR, 2024, pp. 17117–17126.
[58] Z. Wang, H. Huang, A. Zheng, R. He, Heterogeneous test-time training for multi-modal person re-identification, in: AAAI, 38, 2024, pp. 5850–5858.

Yusong Sheng received his BEng degree in 2022 and is currently pursuing the MEng degree in the School of Artificial Intelligence, Anhui University, Hefei, China. His research interests include Computer Vision, Multi-modal Intelligence and Object Re-identification.

Yuhe Ding is currently a PhD student in the School of Computer Science and Technology, Anhui University, Hefei, China. She received her BEng degree from Anhui University in July 2019. From 2020 to 2022, she was a visiting student at CASIA. Her research interests focus on transfer learning, pattern recognition, and computer vision.

Aihua Zheng received BEng degrees and finished Master-Docter combined program in Computer Science and Technology from Anhui University of China in 2006 and 2008, respectively. And received PhD degree in computer science from University of Greenwich of UK in 2012. She visited University of Stirling and Texas State University during June to September in 2013 and during September 2019 to August 2020 respectively. She is currently a Professor and PhD supervisor at the School of Artificial Intelligence, Anhui University. Her main research interests include vision based artificial intelligence and pattern recognition. Especially on person/vehicle re-identification, audio visual computing, and multi-modal intelligence.

Ziqi Liu is a Canadian high school student who will graduate in 2026 and is currently in Grade 11 at Kingsway College in Oshawa, Ontario. Her research interests include Artificial Intelligence and Computer Engineering.

Zi Wang received his BEng degree and completed the Master-doctor combined program in the School of Computer Science and Technology at Anhui University. He currently works at the School of Biomedical Engineering at Anhui Medical University. He is primarily engaged in research on computer vision, medical image processing, and multi-modal learning.

Jin Tang received the BEng degree in automation and the PhD degree in Computer Science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is currently a Professor and PhD supervisor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision, pattern recognition, and machine learning.