MAPPING AND MATCHING JOINT LEARNING FOR UNPAIRED UNSUPERVISED VISIBLE-INFRARED PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

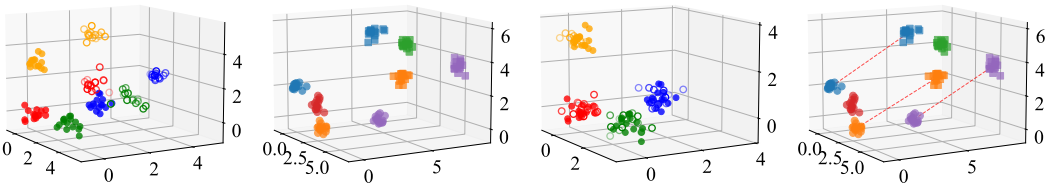
Unsupervised visible-infrared person re-identification (UVI-ReID) aims to associate visible and infrared images of the same identities without relying on manual annotations. Existing approaches typically rely on the unrealistic assumption that the identities between the two modalities are completely overlapped. To advance the development of UVI-ReID, recent studies begin to explore the unpaired UVI-ReID task and introduce feature mapping strategies. Although these studies achieve certain progress, they overlook two critical factors: the identity consistency of the mapped features and the effectiveness of inter-modality overlapped identities (i.e., real cross-modality positive pairs). To this end, we propose a mapping and matching joint learning method (MMJL). First, we introduce an attention-guided feature mapping module, which reconstructs and weights features based on cross-modality attention weights to generate mapped features while preserving intra-identity consistency. Then, we introduce a cross-modality positive matching module, which identifies real cross-modality positive clusters based on the consistency between two cross-modality matching results and the similarity between the intra-modality mapped features and the target features. Moreover, we introduce multiple contrastive losses to simultaneously ensure the identity discriminability and modality invariance of person features. Extensive experiments on two unpaired visible-infrared datasets demonstrate the superiority of the proposed method and the effectiveness of each component.

1 INTRODUCTION

Person re-identification (ReID) (Ye et al., 2021b; Pang et al., 2023; Sun et al., 2024; Ye et al., 2025) aims to retrieve images with specific identities from person images captured by different cameras. Existing surveillance cameras typically capture visible images under sufficient illumination conditions and infrared images under low-light conditions, such as at night. Accordingly, researchers propose a large number of visible-infrared ReID (VI-ReID) methods (Feng et al., 2019a; Lu et al., 2020; Ye et al., 2021a; Yu et al., 2023; Ren & Zhang, 2024; Zhang et al., 2025), which focus on matching images of the same identity across visible and infrared modalities. To reduce the cost of manual annotations, a large number of unsupervised VI-ReID (UVI-ReID) methods (Yang et al., 2022; Wu & Ye, 2023; Yang et al., 2023; 2024; Shi et al., 2024a; Pang et al., 2025a;b) have emerged in recent years. These UVI-ReID methods typically first cluster image features from different modalities separately to generate image clusters within each modality. Subsequently, they enforce the matching of cross-modality positive cluster pairs based on the similarity between cross-modality cluster centroids, thereby obtaining inter-modality pseudo-labels. In this paper, we refer to one visible image (feature) and one infrared image (feature) of the same identity (or pseudo-label) as a cross-modality positive pair, which provides data support for modality-invariant learning.

Although existing UVI-ReID methods do not rely on explicit manual identity labels, they typically rely on an implicit identity assumption, i.e., the identity sets of person images in the visible and infrared modalities are completely overlapped. However, since visible and infrared images are usually captured during the daytime and nighttime, respectively, and the persons appearing during the day and night are generally different, the assumption is hardly satisfied in real-world unsupervised scenarios. Therefore, recent studies begin to explore the unpaired UVI-ReID task (Yao et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) MCL (b) Unconnected identities (c) AFM (d) Connected identities

Figure 1: Comparison between existing methods and our method. (a) and (c) show the feature mapping results of the existing method MCL and the proposed AFM, respectively. (b) and (d) illustrate the feature distributions of some identities that are overlapped and non-overlapped between modalities. The inter-modality overlapped identities are connected by red dashed lines in (d). Solid circles and squares represent real infrared and visible features, respectively. Hollow circles denote the features mapped from the visible modality to the infrared modality. Different colors of the markers (circles and squares) indicate different identities.

2025), which not only prohibits the use of manual identity labels but also assumes that the inter-modality identity sets are not completely overlapped. For example, MCL (Yao et al., 2025) performs cross-modality mapping for each image feature based on the modality-specific mean and variance. The mapped feature and the original feature are regarded as a (generated) cross-modality positive pair. During optimization, MCL increases the similarity between these cross-modality positive pairs, thereby guiding the model to extract modality-invariant and identity-related image features.

Although MCL (Yao et al., 2025) effectively promotes modality-invariant learning, it overlooks two critical factors. On the one hand, the feature mapping process only relies on the mean and variance, which makes it difficult to preserve identity consistency, i.e., the similarity relationships between the mapped features and the features within the target modality. Theoretically, features of the same identity should maintain certain similarity (Yang et al., 2024; Pang et al., 2025a). However, in Fig. 1a, the mapped features in the infrared modality exhibit lower similarity with the real infrared features of the same identity. This significantly affects the subsequent optimization direction. On the other hand, although the identity sets between the visible and infrared modalities are not completely overlapped, there still exist partially overlapped identities between the two modalities, as shown in Fig. 1b. They have the potential to further promote modality-invariant learning. However, MCL only focuses on the generated cross-modality positive pairs while ignoring the inter-modality overlapped identities, which prevents the model from obtaining more diverse supervisory signals.

To address the above issues, we design a mapping and matching joint learning method (MMJL). As shown in Fig. 1c and Fig. 1d, MMJL aims to preserve the intra-identity consistency of mapped features and exploit inter-modality overlapped identities. First, we introduce an attention-guided feature mapping module (AFM). Specifically, AFM first calculates the attention weights between image features and all cluster centroids within the target modality. Then, it obtains reconstructed features based on the attention weights and the cluster centroids of the target modality. Finally, it adaptively fuses the features before and after reconstruction through weighted aggregation, thereby generating mapped features while preserving intra-identity consistency. We regard the features before and after mapping as a generated cross-modality positive pair. To obtain real cross-modality positive pairs, we introduce a cross-modality positive matching module (CPM). Specifically, CPM first assigns the mapped features the same pseudo-labels as the original features, thereby obtaining mapped image clusters within both modalities. Then, it performs matching between real visible clusters and infrared clusters, as well as between mapped visible clusters and infrared clusters. Each matched pair in the consistent matching results is regarded as a cross-modality positive cluster. Within each cross-modality positive cluster, any two images from the two corresponding clusters are considered a real cross-modality positive pair. Furthermore, to avoid discarding some highly similar inter-modality positive clusters, we further select additional real cross-modality positive clusters from the inconsistent matching results based on the intra-modality similarity between clusters. During the optimization stage, in addition to introducing an intra-modality cluster contrastive loss, we also incorporate an inter-modality mapping contrastive loss, an inter-modality matching contrastive loss, and an inter-modality discriminative contrastive loss, so as to simultaneously leverage both the generated and real cross-modality positive pairs to promote modality-invariant identity learning.

The contributions of this work are summarized as follows:

- We propose an unpaired UVI-ReID method, termed MMJL, which aims to preserve intra-identity consistency after cross-modality feature mapping and exploit inter-modality overlapped identities, thereby promoting modality-invariant learning.
- We abandon the traditional feature mapping scheme based on mean and variance and instead introduce the AFM, which considers the inter-modality identity similarity relationships during the feature mapping process, thereby preserving the intra-identity consistency.
- We introduce the CPM, which aims to identify real cross-modality overlapping identities by leveraging the consistency of cross-modality matching results and the similarity of intra-modality features, thereby facilitating modality-invariant feature learning.
- We introduce loss functions for the generated and real cross-modality positive pairs, respectively, to jointly promote modality-invariant learning. Extensive experiments on unpaired VI-ReID datasets validate the effectiveness of the proposed method.

2 Related Work

Given a visible image, VI-ReID aims to retrieve images with the same identity from the infrared modality, and vice versa. Early VI-ReID studies typically rely on manually annotated identity labels and are therefore categorized as supervised VI-ReID (SVI-ReID) methods. SVI-ReID methods focus on designing sophisticated network architectures and multi-dimensional optimization strategies to fully exploit manual identity labels for enhancing modality-invariant learning. For example, early studies (Feng et al., 2019b; Ye et al., 2021b) generally employ dual-branch encoders for feature extraction, where the front part of the encoder contains two modality-specific branches with independent parameters to handle modality discrepancies, while the back part shares parameters between the two modalities to learn modality-shared features. From the perspective of optimization, PMT (Lu et al., 2023) introduces the modality-shared enhancement loss and discriminative center loss during optimization to explore reliable identity features and enhance their discriminability. IDKL (Ren & Zhang, 2024) aims to fully exploit identity information from modality-specific features during optimization and enhances the identity discriminability of modality-shared features through an alignment loss. CSDN (Yu et al., 2025) and DiVE (Dai et al., 2025) introduce the CLIP architecture (Li et al., 2023) to leverage textual semantics for enhancing identity-related learning.

To reduce the cost of manual annotations, existing UVI-ReID methods typically employ clustering and matching algorithms to generate pseudo-labels as substitutes for manually annotated identity labels. In early studies, ADCA (Yang et al., 2022) and PGM (Wu & Ye, 2023) first perform clustering within each modality to generate intra-modality pseudo-labels, and then conduct matching between the two modalities to obtain inter-modality pseudo-labels. This strategy of "intra-modality clustering and inter-modality matching" also provides an effective paradigm for subsequent studies. In later research, SDCL (Yang et al., 2024) and TokenMatcher (Wang et al., 2025) exploit semantic information from vision transformers (ViT) (Dosovitskiy et al., 2021; He et al., 2021) to enhance inter-modality matching. PCLHD (Shi et al., 2024b) introduces hard prototype and dynamic prototype contrastive learning strategies to enhance the optimization strategy.

Although existing UVI-ReID methods theoretically reduce the cost of manual annotations, they still rely on the assumption that the identity sets between visible and infrared modalities are completely overlapped. However, since the identity sets of people appearing during the daytime and nighttime are usually not completely overlapped, this assumption is difficult to hold in real-world scenarios. To address this issue, Yao et al. begin to explore the unpaired UVI-ReID task and propose the MCL method (Yao et al., 2025). Specifically, MCL first introduces a feature mapping strategy to generate cross-modality correspondences for each image feature, which are further constructed into cross-modality positive pairs. During optimization, MCL promotes modality-invariant learning by increasing the consistency between these generated cross-modality positive pairs. Although this method achieves certain performance improvements, it ignores the identity consistency of mapped features and the role of real cross-modality positive pairs, thereby reducing the diversity of supervisory signals. Therefore, this paper aims to simultaneously obtain and utilize reliable generated and real cross-modality positive pairs, enriching the diversity of cross-modality supervisory signals and further promoting modality-invariant feature learning.

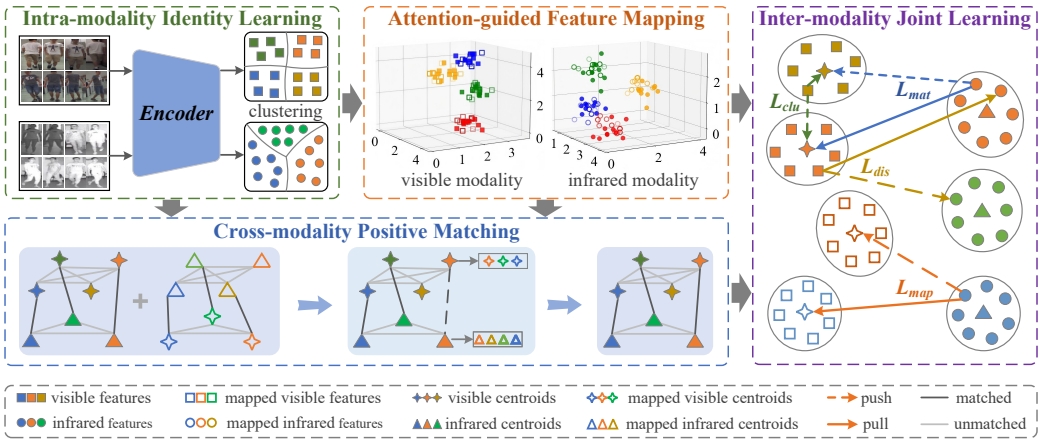


Figure 2: Illustration of the proposed mapping and matching joint learning method (MMJL).

3 The Proposed Method

In the unpaired UVI-ReID task, our goal is to train an image encoder based on unlabeled visible image set $\{x_i^{vis}\}_{i=1}^{N_{vis}}$ and infrared image set $\{x_i^{ir}\}_{i=1}^{N_{ir}}$, such that it can extract similar features for visible and infrared images of the same identity. Here, N_{vis} and N_{ir} denote the numbers of visible and infrared images, respectively, and the identity sets of $\{x_i^{vis}\}_{i=1}^{N_{vis}}$ and $\{x_i^{ir}\}_{i=1}^{N_{ir}}$ are not completely overlapped. The proposed MMJL is illustrated in Fig. 2. MMJL first introduces intra-modality identity learning, which guides the encoder to learn identity-related image features by minimizing the intra-modality cluster contrastive loss L_{clu} based on the clustering results within each modality. Then, MMJL iteratively performs attention-guided feature mapping, cross-modality positive matching, and inter-modality joint learning to promote modality-invariant learning. In inter-modality joint learning, MMJL introduces the intra-modality cluster contrastive loss L_{clu}, inter-modality mapping contrastive loss L_{map}, inter-modality matching contrastive loss L_{mat}, and inter-modality discriminative contrastive loss L_{dis} to jointly guide the optimization of the encoder.

3.1 Intra-Modality Identity Learning

At the beginning of each epoch, MMJL first employs the image encoder to extract features for all images in the visible and infrared modalities, obtaining the visible feature set $\{f_i^{vis}\}_{i=1}^{N_{vis}}$ and infrared feature set $\{f_i^{ir}\}_{i=1}^{N_{ir}}$, respectively. Then, it performs clustering on the feature sets within each modality and calculates the centroid of each cluster. For example, the k-th cluster centroids of the visible and infrared modalities are defined as:

$$c_k^{vis} = \frac{1}{n_k^{vis}} \sum_{i=1}^{n_k^{vis}} f_i^{vis}, c_k^{ir} = \frac{1}{n_k^{ir}} \sum_{i=1}^{n_k^{ir}} f_i^{ir} \tag{1}$$

where n_k^{vis} and n_k^{ir} denote the numbers of features within the corresponding clusters, respectively. Then, we introduce the intra-modality cluster contrastive loss within each modality. For example, the combination of intra-modality cluster contrastive losses for an arbitrary visible feature f_i^{vis} and an arbitrary infrared feature f_j^{ir} is defined as:

$$L_{clu} = \log \frac{\sum_{u=1}^{|C_{vis}|} \exp(f_i^{vis} \cdot {c_u^{vis}}^{\mathrm{T}}/\tau)}{\exp(f_i^{vis} \cdot {c_p^{vis}}^{\mathrm{T}}/\tau)} + \log \frac{\sum_{u=1}^{|C_{ir}|} \exp(f_j^{ir} \cdot {c_u^{ir}}^{\mathrm{T}}/\tau)}{\exp(f_i^{ir} \cdot {c_q^{ir}}^{\mathrm{T}}/\tau)} \tag{2}$$

where c_p^{vis} and c_q^{ir} denote the centroids of the clusters to which f_i^{vis} and f_j^{ir} belong, respectively. C_{vis} and C_{ir} represent the numbers of clusters in the visible and infrared modalities during the cur-

rent epoch, respectively, and τ denotes the temperature hyperparameter. The above intra-modality clustering, centroid calculation, and optimization processes are performed iteratively.

3.2 Attention-Guided Feature Mapping

Although intra-modality identity learning guides the encoder to extract similar identity features for images of the same identity within the same modality, it does not consider the inter-modality identity feature consistency. To obtain sufficient cross-modality positive pairs in the unpaired UVI-ReID task, we introduce AFM. Specifically, AFM first follows Eq. 1 to calculate the centroid of each cluster within each modality and regards the centroids as the identity representatives of the clusters. Then, it obtains attention weights based on the similarities between image features and all centroids within the target (another) modality, and generates reconstructed features based on the attention weights and the cluster centroids of the target modality. For example, the reconstructed feature corresponding to an arbitrary infrared feature f_i^{ir} is calculated as:

$$\tilde{f}_i^{vis} = \sum_{k=1}^{|C_{vis}|} \left[\frac{\exp\left(f_i^{ir} \cdot {c_k^{vis}}^{\mathrm{T}}/\tau\right)}{\sum_{u=1}^{|C_{vis}|} \exp\left(f_i^{ir} \cdot {c_u^{vis}}^{\mathrm{T}}/\tau\right)} c_k^{vis} \right] \tag{3}$$

where c_k^{vis} denotes the k-th cluster centroid in the visible modality, and C_{vis} represents the number of cluster centroids within the visible modality. Considering that there may not exist features with the same real identity as f_i^{ir} in the visible modality, we perform weighted fusion of the features before and after reconstruction to preserve identity consistency before and after mapping, thereby obtaining the final mapped feature. For example, the mapped feature corresponding to the infrared feature f_i^{ir} is calculated as:

$$\hat{f}_i^{vis} = \tilde{f}_i^{vis} + \min\left(1, \frac{||f_i^{ir} - \tilde{f}_i^{vis}||_2}{||f_i^{ir}||_2 + \varepsilon}\right)\left(f_i^{ir} - \tilde{f}_i^{vis}\right) \tag{4}$$

where ε is a small value used to maintain computational stability. For any visible feature f_i^{vis}, we can obtain the mapped feature $\hat{f}_i^{ir}$ following the same strategy. Finally, we regard the features before and after mapping (e.g., f_i^{ir} and $\hat{f}_i^{vis}$) as a generated cross-modality positive pair.

3.3 Cross-Modality Positive Matching

In the unpaired UVI-ReID task, the identity sets of $\{x_i^{vis}\}_{i=1}^{N_{vis}}$ and $\{x_i^{ir}\}_{i=1}^{N_{ir}}$ partially overlap, and these overlapped identities can usually provide additional supervisory signals for model optimization. Therefore, we introduce CPM to exploit these inter-modality overlapped identities. CPM first assigns the mapped features generated by AFM with the same pseudo-labels as the original features, thereby obtaining mapped image clusters within each modality. Then, CPM follows Eq. 1 to calculate the centroid of each mapped cluster within each modality. For example, the k-th mapped cluster centroids of the visible and infrared modalities are defined as:

$$\hat{c}_k^{vis} = \frac{1}{n_k^{ir}} \sum_{i=1}^{n_k^{ir}} \hat{f}_i^{vis}, \hat{c}_k^{ir} = \frac{1}{n_k^{vis}} \sum_{i=1}^{n_k^{vis}} \hat{f}_i^{ir} \tag{5}$$

Then, CPM calculates the assignment cost between real visible and infrared clusters. For example, the assignment cost between the a-th visible cluster and the b-th infrared cluster is calculated as:

$$A\left(c_a^{vis}, c_b^{ir}\right) = \exp\left(-\frac{c_a^{vis} \cdot c_b^{ir}}{||c_a^{vis}|| \times ||c_b^{ir}||}\right) \tag{6}$$

where c_a^{vis} and c_b^{ir} denote the centroids of the a-th visible cluster and the b-th infrared cluster, respectively. Similarly, we can calculate the assignment cost between any visible mapped cluster

and any infrared mapped cluster. Then, CPM introduces basic graph matching (Wu & Ye, 2023) between real visible clusters and infrared clusters, and between mapped visible clusters and infrared clusters, respectively, to obtain two sets of matching results. Subsequently, CPM determines the consistency between these two sets of matching results and regards each consistent matched pair as a real cross-modality positive cluster.

Furthermore, to avoid discarding highly similar cross-modality positive clusters that may originate from the same identity, CPM further selects additional real cross-modality positive clusters from the inconsistent matching results based on inter-cluster similarity. For example, suppose that the i-th cluster centroid c_i^{ir} in the infrared modality is matched with the j-th cluster centroid c_j^{vis} in the visible modality, whereas the visible mapped cluster centroid $\hat{c}_i^{vis}$ corresponding to c_i^{ir} is not matched with the infrared mapped cluster centroid $\hat{c}_j^{ir}$ corresponding to c_j^{vis}. In this case, we perform the following two evaluations: (1) We sort all visible mapped cluster centroids in descending order according to their cosine similarities with c_j^{vis}, and determine whether $\hat{c}_i^{vis}$ is within the top h proportion of the sorted sequence. (2) We sort all infrared mapped cluster centroids in descending order according to their cosine similarities with c_i^{ir}, and determine whether $\hat{c}_j^{ir}$ is within the top h proportion of the sorted sequence. If both conditions are satisfied, it indicates that c_i^{ir} and c_j^{vis} indeed have high similarity. Therefore, we also regard the two clusters corresponding to c_i^{ir} and c_j^{vis} as a real cross-modality positive cluster pair. Any two images (features) from the two corresponding clusters within a cross-modality positive cluster are considered a real cross-modality positive pair.

3.4 Inter-Modality Joint Learning

To promote modality-invariant learning, we simultaneously exploit the generated and real cross-modality positive pairs, and introduce the inter-modality mapping contrastive loss, inter-modality matching contrastive loss, and inter-modality discriminative contrastive loss.

The inter-modality mapping contrastive loss and inter-modality matching contrastive loss focus on improving the consistency between the generated and real cross-modality positive pairs, respectively. For example, the combination of inter-modality mapping contrastive losses for an arbitrary visible feature f_i^{vis} and an arbitrary infrared feature f_j^{ir} is defined as:

$$L_{map} = \log \frac{\sum_{u=1}^{|C_{vis}|} \exp(f_i^{vis} \cdot \hat{c}_u^{ir\,\mathrm{T}}/\tau)}{\exp(f_i^{vis} \cdot \hat{c}_p^{ir\,\mathrm{T}}/\tau)} + \log \frac{\sum_{u=1}^{|C_{ir}|} \exp(f_j^{ir} \cdot \hat{c}_u^{vis\,\mathrm{T}}/\tau)}{\exp(f_i^{ir} \cdot \hat{c}_q^{vis\,\mathrm{T}}/\tau)} \tag{7}$$

where $\hat{c}_u^{ir}$ and $\hat{c}_u^{vis}$ denote the centroids of the mapped clusters corresponding to the clusters to which f_i^{vis} and f_j^{ir} belong, respectively. For the real cross-modality positive pairs obtained from CPM, the combination of inter-modality matching contrastive losses for an arbitrary visible feature f_i^{vis} and an arbitrary infrared feature f_j^{ir} is defined as:

$$L_{mat} = \log \frac{\sum_{u=1}^{|C_{ir}|} \exp(f_i^{vis} \cdot c_u^{ir\,\mathrm{T}}/\tau)}{\exp(f_i^{vis} \cdot c_v^{ir\,\mathrm{T}}/\tau)} + \log \frac{\sum_{u=1}^{|C_{vis}|} \exp(f_j^{ir} \cdot c_u^{vis\,\mathrm{T}}/\tau)}{\exp(f_i^{ir} \cdot c_w^{vis\,\mathrm{T}}/\tau)} \tag{8}$$

where c_v^{ir} and c_w^{vis} denote the centroids of the cross-modality positive clusters corresponding to the clusters to which f_i^{vis} and f_j^{ir} belong, respectively.

Although the inter-modality matching contrastive loss can improve the cross-modality consistency of some potential inter-modality overlapped identities, it does not fully consider the similarity relationships among inter-modality non-overlapped identities. Theoretically, inter-modality non-overlapped identities should maintain sufficient discriminability and distinctiveness. Therefore, we introduce the inter-modality discriminative contrastive loss. For any image feature f_i^{vis} (f_i^{ir}) that has real cross-modality positive features through CPM, we search for the cross-modality positive feature with the minimum similarity with f_i^{vis} (f_i^{ir}), denoted as f_{min}^{ir} (f_{min}^{vis}), within the corresponding minibatch. Meanwhile, we define the sets of visible and infrared features without real cross-modality positive features obtained by CPM in the minibatch as U_i^{vis} and U_i^{ir}, respectively. Therefore, the combination of inter-modality discriminative contrastive losses for any visible feature

f_i^{vis} and any infrared feature f_j^{ir} that have real cross-modality positive features through CPM in the minibatch is defined as:

$$L_{dis} = \log \frac{\sum_{f_u^{ir} \in U_i^{ir} \cup f_{\min}^{ir}} \exp(f_i^{vis} \cdot {f_u^{ir}}^{\mathrm{T}}/\tau)}{\exp(f_i^{vis} \cdot {f_{\min}^{ir}}^{\mathrm{T}}/\tau)} + \log \frac{\sum_{f_u^{vis} \in U_i^{vis} \cup f_{\min}^{vis}} \exp(f_j^{ir} \cdot {f_u^{vis}}^{\mathrm{T}}/\tau)}{\exp(f_i^{ir} \cdot {f_{\min}^{vis}}^{\mathrm{T}}/\tau)} \quad (9)$$

In summary, the overall loss function of MMJL is defined as:

$$L_{mmjl} = L_{clu} + L_{map} + L_{mat} + \lambda_{dis} L_{dis} \quad (10)$$

where λ_{dis} is the weighting hyperparameter of L_{dis}.

4 Experiment

4.1 Datasets and Implementation Details

The datasets used in our experiments include unpaired SYSU-MM01 (Wu et al., 2017) and RegDB (Nguyen et al., 2017). To simulate the unpaired UVI-ReID scenario, we process the training sets of the above datasets separately. Specifically, we set the identity overlap ratios between modalities to 0.25, 0.50, and 0.75, respectively. We use Rank-1, mAP and mINP as the performance evaluation metrics. Detailed information on the datasets is provided in the Appendix B.

In this paper, all images are resized to 288×144 pixels. Within each modality, we employ DBSCAN (Ester et al., 1996) to cluster image features and set the minimum number of samples to 4. For the unpaired SYSU-MM01 and RegDB datasets, the distance thresholds of DBSCAN are set to 0.6 and 0.3, respectively. For the image encoder, we adopt AGW (Ye et al., 2021b) with a ResNet-50 (He et al., 2016) backbone as the network architecture and initialize it with the weights pre-trained on ImageNet (Deng et al., 2009). The training of the encoder consists of two stages. Intra-modality identity learning is first performed for 50 epochs, after which AFM, CPM, and inter-modality joint learning are iteratively executed for another 50 epochs. During the training stage, random cropping, random flipping, and random erasing (Zhong et al., 2020) are employed for data augmentation. In addition, channel augmentation (Ye et al., 2021a) is applied to visible images. For CPM, we set the ratio threshold to $h = 0.10$. For encoder optimization, we adopt the Adam optimizer (Kingma & Ba, 2015) and set the initial learning rate to 0.00035. In each minibatch, we randomly sample 8 clusters from each modality and randomly sample 16 images from each cluster. Meanwhile, we ensure that each modality in every minibatch contains at least one cluster with a cross-modality positive cluster. The temperature hyperparameter τ is set to 0.05. We set the small value ε to 10^{-6}. For the weighting hyperparameter, we set $\lambda_{dis} = 0.50$. The analyses of h and λ_{dis} are provided in the Appendix D.

4.2 Comparison with the State-of-the-art Methods

We evaluate the proposed MMJL on the unpaired SYSU-MM01 and RegDB datasets. The compared methods include PGM (Wu & Ye, 2023), PCLHD (Shi et al., 2024b), PCAL (Yang et al., 2025), MCL (Yao et al., 2025), and BMIL (Wang et al., 2026). The performances of the compared methods in Table 1 are all reproduced by us under the testing scenarios established in this paper. MCL and BMIL have achieved better performance in existing methods. As shown in Table 1, the proposed MMJL method outperforms existing methods under all three overlap ratios of 0.25, 0.50, and 0.75. For example, on SYSU-MM01 with an overlap ratio of 0.50, MMJL achieves 4.83% and 5.10% higher Rank-1 accuracy than BMIL under the All Search and Indoor Search settings, respectively. Furthermore, we observe that the advantage of MMJL over existing methods becomes increasingly significant as the overlap ratio decreases. For example, on SYSU-MM01 with an overlap ratio of 0.75, MMJL achieves a 2.02% higher Rank-1 accuracy than BMIL under the All Search setting. On SYSU-MM01 with an overlap ratio of 0.25, MMJL achieves a 9.43% higher Rank-1 accuracy than BMIL under the All Search setting. This is mainly because MMJL not only mines inter-modality overlapped identities while ensuring the high reliability of cross-modality positive pairs, but also preserves the intra-identity consistency between mapped features and overlapped identity features

Table 1: Comparison with existing methods on the unpaired SYSU-MM01 and RegDB datasets. The top two performances in each setting are highlighted in bold and underline, respectively.

Overlap Ratios	Methods	SYSU-MM01						RegDB					
		All Search			Indoor Search			Visible to Infrared			Infrared to Visible		
		Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP
0.25	PGM	35.47	34.16	21.44	40.17	49.47	45.18	43.83	41.19	30.04	44.13	40.75	27.21
	PCLHD	39.99	37.65	23.90	43.36	51.53	46.78	56.99	52.84	39.27	57.02	52.90	38.08
	PCAL	38.69	37.97	22.06	41.79	50.02	46.25	53.04	50.39	36.29	55.87	50.27	36.75
	MCL	43.89	44.07	30.12	50.23	58.67	55.25	70.36	63.29	47.66	70.03	62.87	46.71
	BMIL	42.93	42.39	28.98	48.95	57.31	52.98	68.81	62.50	48.53	68.82	61.78	45.48
	MMJL	**52.36**	**51.97**	**38.05**	**58.23**	**66.59**	**62.14**	**77.66**	**68.91**	**51.46**	**76.28**	**66.57**	**49.55**
0.50	PGM	40.53	37.99	23.46	43.40	51.67	47.16	56.50	53.21	40.27	56.02	52.85	39.67
	PCLHD	45.32	43.71	30.05	50.20	58.77	54.48	68.83	64.03	52.14	67.90	62.50	48.07
	PCAL	43.10	40.64	27.65	46.75	55.36	50.89	62.76	59.04	46.99	61.66	57.59	44.90
	MCL	50.06	48.14	34.26	57.73	64.20	60.89	79.26	72.57	60.25	78.67	72.10	56.92
	BMIL	50.69	49.23	35.30	58.12	65.46	61.43	80.66	73.99	60.78	80.45	73.49	58.14
	MMJL	**55.52**	**55.07**	**39.65**	**63.22**	**70.04**	**65.17**	**85.67**	**77.28**	**62.35**	**86.29**	**77.03**	**59.94**
0.75	PGM	43.92	42.52	28.67	48.72	56.33	52.17	64.37	59.30	44.40	63.89	58.66	43.34
	PCLHD	56.81	53.52	38.76	60.42	66.61	62.02	75.02	70.33	58.15	74.85	69.17	54.45
	PCAL	53.25	51.80	36.63	57.29	63.05	59.63	72.63	68.22	55.86	72.10	67.04	52.08
	MCL	56.87	53.69	38.60	62.40	68.56	64.95	85.29	79.30	66.27	87.13	78.66	63.27
	BMIL	57.61	55.20	39.89	62.70	68.85	65.12	88.78	82.08	69.17	89.06	81.92	66.90
	MMJL	**59.63**	**58.94**	**46.21**	**65.33**	**71.89**	**68.67**	**91.26**	**85.04**	**70.16**	**90.15**	**82.07**	**67.25**

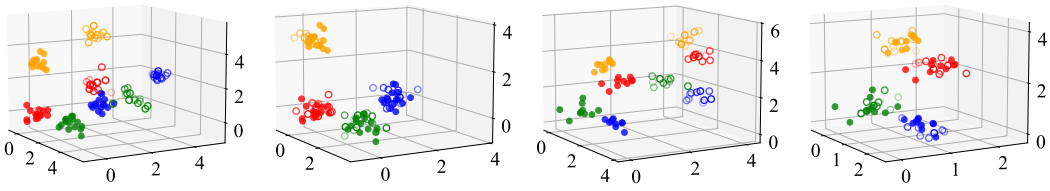
Table 2: Ablation study (overlap ratio is 0.5) on the unpaired SYSU-MM01 and RegDB datasets.

Methods	mapping	CPM	L_{dis}	SYSU-MM01 (All Search)			SYSU-MM01 (Indoor Search)			RegDB (Visible to Infrared)		
				Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP
Baseline				19.27	19.36	10.69	23.25	32.59	29.28	11.56	13.60	9.78
M1	MCL			50.06	48.14	34.26	57.73	64.20	60.89	79.26	72.57	60.25
M2	AFM			51.37	49.55	36.04	59.27	65.69	61.37	80.79	74.20	60.76
M3	MCL	✓		51.56	50.20	36.23	59.79	66.05	61.86	81.25	74.96	60.61
M4	AFM	✓		54.49	53.87	38.26	62.06	68.59	63.88	84.02	75.96	61.94
MMJL	AFM	✓	✓	55.52	55.07	39.65	63.22	70.04	65.17	85.67	77.28	62.35

in the target modality by considering the inter-modality identity similarity relationships. The above experimental results validate the superiority and robustness of MMJL.

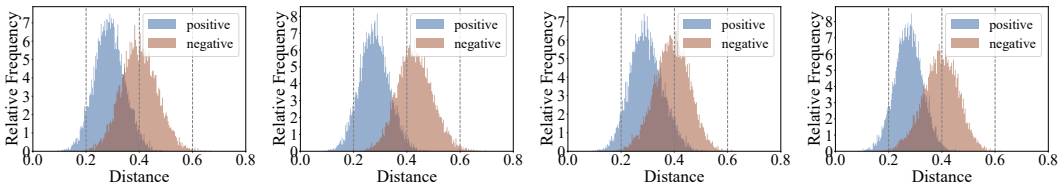
4.3 Ablation Study

Effectiveness of AFM. As shown in Table 2, M1 and M2 introduce MCL (Yao et al., 2025) and AFM into the baseline for cross-modality feature mapping, respectively, and utilize the mapped features to assist model optimization. We observe that both M1 and M2 achieve significant performance improvements over the baseline. For example, compared with the baseline, M1 improves the Rank-1 accuracy by 30.79%, 34.48%, and 67.70% on SYSU-MM01 (All Search), SYSU-MM01 (Indoor Search), and RegDB (Visible to Infrared), respectively. This verifies the effectiveness of cross-modality feature mapping in the unpaired UVI-ReID task. Furthermore, we observe that M2 outperforms M1. Specifically, M2 improves the Rank-1 accuracy over M1 by 1.31%, 1.54%, and 1.53% under the above three testing settings, respectively. To further investigate the reasons behind this performance improvement, we visualize the cross-modality mapped features generated by MCL and AFM, respectively. As shown in Fig. 3, the mapped infrared features generated by AFM are closer to the features with the same real identity within the infrared modality, whereas the mapped features generated by MCL generally fail to maintain a high degree of compactness with the features of the same identity in the target modality. This verifies that AFM can preserve the inter-modality identity consistency, thereby enhancing inter-modality identity learning.



(a) MCL(SYSU-MM01) (b) AFM(SYSU-MM01) (c) MCL(RegDB) (d) AFM(RegDB)

Figure 3: Feature mapping results of MCL and AFM. Solid and hollow circles represent the real infrared features and the features mapped from visible features to the infrared modality (i.e., mapped infrared features), respectively. Different colors of circles indicate different identities.



(a) M4(SYSU-MM01) (b) MMJL(SYSU-MM01) (c) M4(RegDB) (d) MMJL(RegDB)

Figure 4: Feature distance distributions of inter-modality positive pairs and inter-modality negative pairs in M4 and MMJL.

Effectiveness of CPM. M3 and M4 introduce CPM based on M1 and M2, respectively, to mine inter-modality overlapped identities, thereby promoting model optimization. As shown in Table 2, M3 and M4 outperform M1 and M2, respectively. For example, compared with M1, M3 improves the Rank-1 accuracy by 1.50%, 2.06%, and 1.99% on SYSU-MM01 (All Search), SYSU-MM01 (Indoor Search), and RegDB (Visible to Infrared), respectively. Compared with M2, M4 improves the Rank-1 accuracy by 3.12%, 2.79%, and 3.23% under the above three testing settings, respectively. The above experimental results demonstrate that CPM can enrich cross-modality supervisory signals by mining inter-modality overlapped identities, thereby effectively improving the model's cross-modality recognition performance.

Effectiveness of L_{dis}. MMJL introduces L_{dis} based on M4 to enhance the discriminability of inter-modality non-overlapped identities. As shown in Table 2, compared with M4, MMJL improves the Rank-1 accuracy by 1.03%, 1.16%, and 1.65% on SYSU-MM01 (All Search), SYSU-MM01 (Indoor Search), and RegDB (Visible to Infrared), respectively. To further evaluate the effectiveness of L_{dis}, we visualize the feature distance distributions of inter-modality positive pairs and inter-modality negative pairs generated by M4 and MMJL, respectively. As shown in Fig. 4, after introducing L_{dis}, MMJL significantly increases the feature distances of inter-modality negative pairs and separates the distributions of inter-modality positive and negative pairs. The above experimental results demonstrate that L_{dis} can enhance the discriminability of inter-modality non-overlapped identities, thereby improving the recognition performance.

5 Conclusion

In this paper, we propose MMJL and introduce three key components: AFM, CPM, and the inter-modality discriminative contrastive loss. We conduct ablation analyses to verify the effectiveness of these three components. Specifically, AFM preserves the inter-modality identity consistency during the feature mapping process, thereby enhancing inter-modality identity learning; CPM mines inter-modality overlapped identities based on cross-modality matching results and inter-cluster similarity, thereby further promoting modality-invariant learning; and the inter-modality discriminative contrastive loss maintains the high discriminability of inter-modality non-overlapped identities, thereby further improving cross-modality recognition performance. Experimental results on two unpaired visible-infrared datasets demonstrate that the proposed MMJL not only significantly outperforms existing methods but also exhibits stronger robustness in unpaired scenarios.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and to check the correctness of the implementation code. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on the research methodology or experiments, implement methods, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results, and generating synthetic datasets is not applicable to this work. Additionally, we did not use generative AI tools for other research subtasks requiring recommended disclosure. All AI-assisted translations were carefully reviewed and revised by the authors, and the code checked with the assistance of generative AI was independently verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Hao Chen, Benoit Lagadec, and Francois Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14960–14969, 2021.

Wenbo Dai, Lijing Lu, and Zhihang Li. Diffusion-based synthetic data generation for visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11185–11193, 2025.

Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, pp. 1142–1160, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, volume 96, pp. 226–231, 1996.

Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 29:579–590, 2019a.

Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 29:579–590, 2019b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15013–15022, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Siyuan Li, Li Sun, and Qingli Li. CLIP-ReID: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1405–1413, 2023.

Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1835–1843, 2023.

Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13379–13389, 2020.

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.

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.

Zhiqi Pang, Chunyu Wang, Lingling Zhao, and Junjie Wang. Augmented and softened matching for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11100–11109, 2025a.

Zhiqi Pang, Lingling Zhao, Junjie Wang, and Chunyu Wang. Lvlm-driven attribute-aware modeling for visible-infrared person re-identification. *Proceedings of the Neural Information Processing Systems*, 38:1769–1788, 2025b.

Kaijie Ren and Lei Zhang. Implicit discriminative knowledge learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 393–402, 2024.

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 *Proceedings of the European Conference on Computer Vision*, pp. 456–474. Springer, 2024a.

Jiangming Shi, Xiangbo Yin, Yachao Zhang, Yuan Xie, Yanyun Qu, et al. Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification. *Proceedings of the Neural Information Processing Systems*, 37:99715–99734, 2024b.

Xiaoxiao Sun, Yue Yao, Shengjin Wang, Hongdong Li, and Liang Zheng. Alice benchmarks: Connecting real world re-identification with the synthetic. In *International Conference on Learning Representations*, volume 2024, pp. 14424–14442, 2024.

Menglin Wang, Xiaojin Gong, Jiachen Li, and Genlin Ji. Modality-aware bias mitigation and invariance learning for unsupervised visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9975–9983, 2026.

Xiao Wang, Lekai Liu, Bin Yang, Mang Ye, Zheng Wang, and Xin Xu. Tokenmatcher: Diverse tokens matching for unsupervised visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7934–7942, 2025.

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*, pp. 5380–5389, 2017.

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*, pp. 9548–9558, 2023.

Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the ACM International Conference on Multimedia*, pp. 2843–2851, 2022.

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*, pp. 11069–11079, 2023.

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*, pp. 16870–16879, 2024.

Yiming Yang, Weipeng Hu, and Haifeng Hu. Progressive cross-modal association learning for unsupervised visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 20:1290–1304, 2025.

Haoyu Yao, Bin Yang, Wenke Huang, Bo Du, and Mang Ye. Unsupervised visible-infrared person re-identification under unpaired settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11916–11926, 2025.

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*, pp. 13567–13576, 2021a.

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, 2021b.

Mang Ye, Shuoyi Chen, Chenyue Li, Wei-Shi Zheng, David Crandall, and Bo Du. Transformer for object re-identification: A survey. *International Journal of Computer Vision*, 133(5):2410–2440, 2025.

Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11185–11195, 2023.

Xiaoyan Yu, Neng Dong, Liehuang Zhu, Hao Peng, and Dapeng Tao. Clip-driven semantic discovery network for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 2025.

Yukang Zhang, Hanzi Wang, Yang Lu, Yan Yan, and Xuelong Li. Frequency domain nuances mining for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 2025.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13001–13008, 2020.

APPENDIX

A ALGORITHM PROCEDURE

The overall algorithmic procedure of MMJL is presented in Algorithm 1. We set $iters$ to 200 on SYSU-MM01 and to 100 on RegDB, respectively.

Algorithm 1: MMJL

Input: Unlabeled data $\{x_i^{vis}\}_{i=1}^{N_{vis}}$ and $\{x_i^{ir}\}_{i=1}^{N_{ir}}$, image encoder parameterized by θ_E, training $epochs$ and $iters$.

for $i = 1$ **to** $epochs$ **do**
- Extract features of $\{x_i^{vis}\}_{i=1}^{N_{vis}}$ and $\{x_i^{ir}\}_{i=1}^{N_{ir}}$ with θ_E;
- Perform clustering on features within each modality separately;
- Calculate cluster centroids based on $Eq.$ 1;
- **for** $j = 1$ **to** $iters$ **do**
  - Optimize θ_E to minimize the loss defined in $Eq.$ 2;
- **end**

end

for $i = 1$ **to** $epochs$ **do**
- Extract features of $\{x_i^{vis}\}_{i=1}^{N_{vis}}$ and $\{x_i^{ir}\}_{i=1}^{N_{ir}}$ with θ_E;
- Perform clustering on features within each modality separately;
- Calculate cluster centroids based on $Eq.$ 1;
- Perform feature mapping with AFM to obtain generated cross-modality positive pairs;
- Perform matching and similarity computation with CPM to mine real cross-modality positive pairs;
- **for** $j = 1$ **to** $iters$ **do**
  - Optimize θ_E to minimize the loss defined in $Eq.$ 10;
- **end**

end

return *Trained* θ_E.

B DATASET DETAILS

We evaluate the effectiveness of the proposed MMJL on the unpaired SYSU-MM01 (Wu et al., 2017) and RegDB (Nguyen et al., 2017) datasets. In previous UVI-ReID tasks (Yang et al., 2022; Wu & Ye, 2023; Yang et al., 2024; Shi et al., 2024b; Pang et al., 2025b), the training set of SYSU-MM01 contains 22,258 visible images and 11,909 infrared images from 395 identities. The researchers set up All Search and Indoor Search test modes and reported the average performance of 10 random gallery selections. The RegDB dataset contains 8,240 images from 412 identities, where each identity corresponds to 10 visible images and 10 thermal infrared images. Researchers generally randomly select images from 206 identities for training and use the remaining images for testing, then report the average performance over 10 random trials.

To simulate the unpaired UVI-ReID scenario, we process the training sets of the above datasets separately. It should be noted that, since we cannot access the training set of MCL (Yao et al., 2025), we establish the following training setting. For the training set of each processed dataset, we set the inter-modality identity overlap ratios to 0.25, 0.50, and 0.75, respectively. Table 3 presents the numbers of identities contained in the visible and infrared modalities of the unpaired SYSU-MM01 (Wu et al., 2017) and RegDB (Nguyen et al., 2017) training sets under cross-modality identity overlap ratios of 0.25, 0.50, and 0.75, respectively. We report the Rank-1, mAP, and mINP of the encoders trained with training sets under different overlap ratios, while using the same testing set as those of the original SYSU-MM01 and RegDB datasets.

Table 3: Number of identities contained in the visible and infrared modalities of the two datasets.

Overlap Ratios	Unpaired SYSU-MM01		Unpaired RegDB	
	Visible	Infrared	Visible	Infrared
0.25	247	246	129	129
0.50	296	296	154	154
0.75	346	345	180	180

Table 4: Ablation study on the unpaired SYSU-MM01 and RegDB datasets under overlap ratios of 0.25, 0.50, and 0.75.

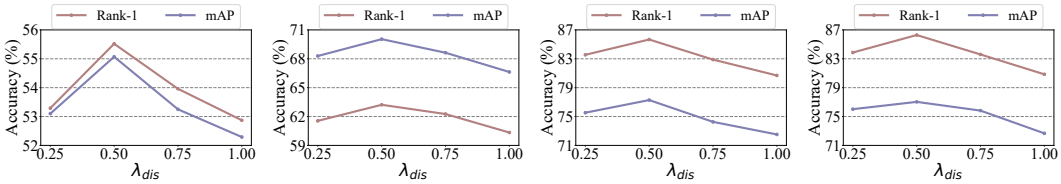
Overlap Ratios	Methods	mapping	CPM	L_{dis}	SYSU-MM01 (All Search)			SYSU-MM01 (Indoor Search)			RegDB (Visible to Infrared)		
					Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP
0.25	Baseline				18.25	18.30	10.03	22.67	31.04	28.03	10.92	13.16	8.93
	M1	MCL			43.89	44.07	30.12	50.23	58.67	55.25	70.36	63.29	47.66
	M2	AFM			45.26	44.83	32.09	52.07	60.44	57.13	71.85	64.76	49.04
	M3	MCL	✓		45.76	45.02	33.25	52.80	61.25	58.06	72.79	65.27	50.17
	M4	AFM	✓		51.03	51.09	37.14	57.26	65.87	61.05	76.29	67.84	51.09
	MMJL	AFM	✓	✓	52.36	51.97	38.05	58.23	66.59	62.14	77.66	68.91	51.46
0.50	Baseline				19.27	19.36	10.69	23.25	32.59	29.28	11.56	13.60	9.78
	M1	MCL			50.06	48.14	34.26	57.73	64.20	60.89	79.26	72.57	60.25
	M2	AFM			51.37	49.55	36.04	59.27	65.69	61.37	80.79	74.20	60.76
	M3	MCL	✓		51.56	50.20	36.23	59.79	66.05	61.86	81.25	74.96	60.61
	M4	AFM	✓		54.49	53.87	38.26	62.06	68.59	63.88	84.02	75.96	61.94
	MMJL	AFM	✓	✓	55.52	55.07	39.65	63.22	70.04	65.17	85.67	77.28	62.35
0.75	Baseline				19.33	19.55	10.87	23.63	33.87	29.64	12.55	13.99	9.89
	M1	MCL			56.87	53.69	38.60	62.40	68.56	64.95	85.29	79.30	66.27
	M2	AFM			57.63	54.07	39.81	63.05	69.16	65.96	87.57	82.06	67.39
	M3	MCL	✓		57.95	55.86	41.37	63.78	69.23	66.29	88.06	82.67	68.05
	M4	AFM	✓		59.14	57.82	45.75	64.82	71.22	68.03	90.65	84.34	69.57
	MMJL	AFM	✓	✓	59.63	58.94	46.21	65.33	71.89	68.67	91.26	85.04	70.16

C Complete Ablation Results

Table 4 presents the ablation results of the three key components in MMJL (i.e., AFM, CPM, and L_{dis}) under cross-modality identity overlap ratios of 0.25, 0.50, and 0.75. The baseline follows the scheme of existing unsupervised conventional ReID methods (Chen et al., 2021; Dai et al., 2022): all visible and infrared images are directly clustered together, and the corresponding L_{clu} and L_{mat} are introduced to promote intra-modality and inter-modality learning, respectively. M1 and M2 introduce MCL (Yao et al., 2025) and AFM, respectively, based on the baseline for feature mapping, and correspondingly introduce L_{map} to further promote modality-invariant learning. However, they only perform intra-modality clustering and do not employ L_{mat}. M3 and M4 further introduce CPM based on M1 and M2, respectively, to mine inter-modality overlapped identities, and introduce L_{mat} to guide the model to extract modality-invariant features. MMJL further introduces L_{dis} based on M4 to enhance the discriminability of inter-modality non-overlapped identities. In Table 4, we observe that, under all three overlap ratios, the performance of the model consistently improves as these three components are gradually introduced. These experimental results further validate the effectiveness of the above three components.

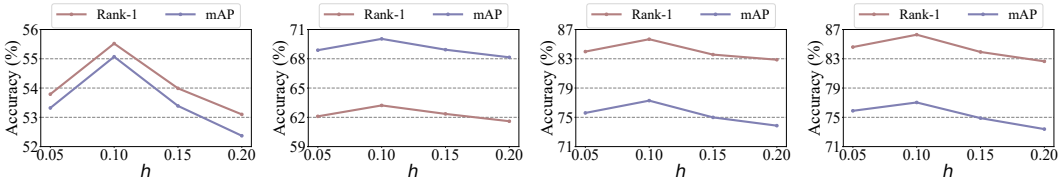
D Parameter Analysis

In this section, we analyze the effects of the weight hyperparameter λ_{dis} of the inter-modality discriminative contrastive loss and the ratio threshold h in CPM on model performance. We report the performance variations of the model under an overlap ratio of 0.5.



(a) SYSU-MM01(All) (b) SYSU-MM01(Indoor) (c) RegDB(V2T) (d) RegDB(T2V)

Figure 5: Sensitivity of model performance to λ_{dis}.



(a) SYSU-MM01(All) (b) SYSU-MM01(Indoor) (c) RegDB(V2T) (d) RegDB(T2V)

Figure 6: Sensitivity of model performance to h.

As shown in Fig. 5, when $\lambda_{dis} = 0.50$, the model achieves relatively superior performance under all four test settings, which verifies the generalizability of λ_{dis}. When λ_{dis} takes excessively large or small values, the performance of the model decreases. This is because an excessively small λ_{dis} reduces the contribution of the inter-modality discriminative contrastive loss to model optimization, thereby affecting the high discriminability of inter-modality non-overlapped identities. In contrast, an excessively large λ_{dis} weakens the contributions of other loss functions, hindering the model from learning identity-related and modality-invariant features. Therefore, we set $\lambda_{dis} = 0.50$.

As shown in Fig. 6, we present the performance variations of the model when the ratio threshold h changes from 0.05 to 0.20. When $h = 0.10$, the model achieves relatively superior performance under all four test settings, which verifies the generalizability of h. When $h = 0.05$, only a few matches from the inconsistent matching results are identified as cross-modality positive pairs, which prevents the model from obtaining sufficient supervisory signals. When h is larger than 0.10, a large number of inconsistent matching results are identified as cross-modality positive pairs, reducing the reliability of cross-modality positive pairs and consequently misleading the optimization direction.