Augmented and Softened Matching for Unsupervised Visible-Infrared Person Re-Identification

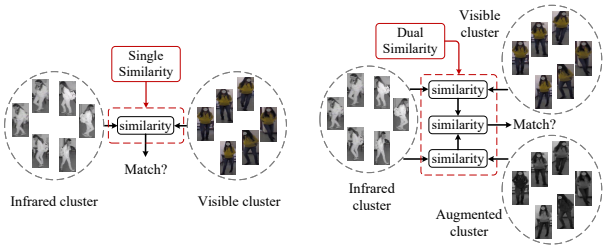
Zhiqi Pang[1] Chunyu Wang[1]* Lingling Zhao[1] Junjie Wang[2]
[1] Harbin Institute of Technology, China
[2] Nanjing Medical University, China
zqpang98@gmail.com, chunyu@hit.edu.cn, zhaoll@hit.edu.cn, junjiehit@163.com

Abstract

Color variations, a key challenge in the unsupervised visible-infrared person re-identification (UVI-ReID) task, have garnered significant attention. While existing UVI-ReID methods have made substantial efforts during the optimization phase to enhance the model's robustness to color variations, they often overlook the impact of color variations on the acquisition of pseudo-labels. To address this, in this paper, we focus on improving the robustness of pseudo-labels to color variations through data augmentation and propose an augmented and softened matching (ASM) method. Specifically, we first develop the cross-modality augmented matching (CAM) module, which performs channel augmentation on visible images to generate augmented images. Then, based on the fusion of the visible-infrared and augmented-infrared centroid similarity matrices, CAM establishes cross-modality correspondences that are robust to color variations. To increase training stability, we design a soft-labels momentum update (SMU) strategy, which converts traditional one-hot labels into soft-labels through momentum updates, thus adapting to CAM. During the optimization phase, we introduce the cross-modality soft contrastive loss and cross-modality hard contrastive loss to promote modality-invariant learning from the perspectives of shared and diversified features, respectively. Extensive experimental results validate the effectiveness of the proposed method, showing that ASM not only outperforms state-of-the-art unsupervised methods but also competes with some supervised methods.

1. Introduction

Visible-infrared person re-identification (VI-ReID) [10, 13, 15, 17, 18, 25, 29] aims to match images of the same person across the visible and infrared modalities. Leveraging various custom model architectures [38, 42] and multiple optimization strategies [40, 41, 45], existing supervised VI-ReID (SVI-ReID) methods have achieved promising performance. However, the heavy reliance on manually annotated identity labels in SVI-ReID severely limits their flexibility. As a result, many researchers have shifted their focus to unsupervised VI-ReID (UVI-ReID) methods [2, 27, 30, 32–36], which typically use automatically generated pseudo-labels to replace the manual identity labels.



(a) Existing methods (b) CAM in ASM

Figure 1. The differences between existing methods and CAM in ASM. (a) Existing methods match the visible clusters with the infrared clusters directly. (b) CAM integrates both visible-infrared and augmented-infrared similarities.

Existing UVI-ReID methods typically first perform intra-modality homogeneous learning, followed by inter-modality heterogeneous learning [30, 32, 36]. The former focuses on enabling the model to acquire preliminary identity recognition capabilities, while the latter emphasizes enhancing the model's robustness to modality gaps, which is the primary focus of existing UVI-ReID methods. In the pseudo-label acquisition stage of heterogeneous learning, current methods first use clustering algorithms [6] to obtain identity pseudo-labels within each modality (i.e., *intra-modality pseudo-labels*), and then apply matching methods to acquire cross-modality correspondences (i.e., *inter-modality pseudo-labels*). During the optimization phase, color variations [41, 43] have emerged as a key factor contributing to the gaps between the visible and infrared modalities, attracting significant attention from researchers. To

*Corresponding author: Chunyu Wang

address this issue, existing methods [30, 32, 35] typically start by applying data augmentation to the training images. For instance, channel augmentation (CA) [41], a widely used data augmentation technique, randomly selects one channel (R, G, or B) of the visible image to replace the other channels. Subsequently, existing methods [30, 32, 35] improve the robustness of the model to color variations by enhancing the feature consistency between the original image and the augmented image generated by CA, both of which share the same pseudo-label.

However, although existing methods have made significant efforts during the optimization phase to enhance the model's robustness to color variations [30, 32, 36], as shown in Fig. 1a, these methods typically perform direct matching between the visible clusters (clusters consisting of visible images) and the infrared clusters (clusters consisting of infrared images) to obtain inter-modality pseudo-labels. This process overlooks the impact of color variations on the acquisition of pseudo-labels. Specifically, considering that CA only alters the color information without changing identity-related information (such as body shape, age, gender, etc.) in the image, the augmented cluster (cluster consisting of images generated by CA) should match the same infrared cluster as the visible cluster. However, according to our statistics, only 51.67% of the matches between the visible and augmented clusters overlap, suggesting that the inter-modality pseudo-labels are not robust to color variations. Low-quality pseudo-labels often misguide the model's optimization direction [30], thus affecting the cross-modality recognition performance. Therefore, enhancing the robustness of pseudo-labels to color variations is an important aspect worth exploring.

To address this, we propose an augmented and softened matching (ASM) method. First, to enhance the robustness of pseudo-labels to color variations, we develop a cross-modality augmented matching (CAM) module within ASM. As shown in Fig. 1b, CAM integrates both visible-infrared and augmented-infrared similarities to obtain cross-modality matches. Specifically, CAM first computes the centroids for all visible, augmented, and infrared clusters. Then, it computes the visible-infrared and augmented-infrared centroid similarity matrices separately. Finally, based on the fusion of these two matrices, it produces matching results that are robust to color variations. Considering that the randomness introduced by data augmentation of CAM can increase the instability of matching results, we design a soft-labels momentum update (SMU) strategy in ASM, which converts traditional one-hot labels into soft-labels through momentum updates, thereby improving training stability. During the optimization phase, we introduce cross-modality soft contrastive loss and cross-modality hard contrastive loss. The former reduces the impact of modality gaps by enhancing the consistency of cross-modality correspondences, while the latter further promotes modality-invariant learning from the perspective of intra-cluster diversity.

The main contributions are summarized as follows:

- We propose an augmented and softened matching (ASM) method for UVI-ReID, which, to our knowledge, is the first work to leverage data augmentation for enhancing the robustness of pseudo-labels against color variations.
- We develop the cross-modality augmented matching (CAM) module, which obtains cross-modality correspondences robust to color variations by fusing visible-infrared and augmented-infrared similarity matrices.
- We not only design the soft-labels momentum update (SMU) strategy to adapt model optimization to the diverse pseudo-labels provided by CAM, but also introduce two cross-modality contrastive losses to promote modality-invariant feature learning.

2. Related Work

2.1. Supervised VI-ReID

The SVI-ReID methods typically rely on manually labeled identity labels to guide the model in learning modality-invariant features. For instance, Zero-Padding [28] preprocesses visible and infrared images separately before feeding the processed images into a single-stream encoder to extract modality-invariant features. Both BDTR [38] and MAC [39] introduce modality-specific layers at the front end of the encoder to adapt to different modalities, and modality-shared layers at the back end to extract modality-invariant features. This dual-stream encoder architecture is also widely adopted in subsequent methods [9, 19, 22, 40]. To promote modality-invariant learning, CAJ [41] introduces auxiliary modalities from the perspective of data augmentation to enhance the model's robustness to color variations. FMCNet [47], on the other hand, incorporates modality-specific information compensation at the feature level to achieve a similar objective. Considering that existing methods typically measure feature distances while neglecting person relationships, SAAI [7] improves the inference process by aligning part features with prototypes. CSDN [46] leverages textual semantics obtained from vision-language models [23, 49] to assist the image encoder in learning modality-invariant features, achieving competitive performance. Although the reliance on manual labels limits the flexibility of SVI-ReID methods, their use of data augmentation and optimization strategies provides important insights for exploring unsupervised methods.

2.2. Unsupervised VI-ReID

Compared to SVI-ReID methods, UVI-ReID methods focus on obtaining identity pseudo-labels to reduce annotation costs. H2H [16], as the first UVI-ReID method, ini-

tially pretrains the encoder on the Market-1501 dataset [48] using manually labeled identity labels, and then fine-tunes the encoder on a cross-modality dataset using the obtained pseudo-labels. Subsequent studies, such as ADCA [32] and CHCR [21], rely solely on pseudo-labels from cross-modality datasets to optimize the model. To address the issue of inconsistent cluster numbers across different modalities, PGM [30] introduces a graph matching strategy to progressively find cross-modality correspondences for all clusters. To leverage the powerful representation capabilities of vision-language models [23, 49], CCLNet [2] first learns a textual semantic for each cluster within each modality and incorporates the textual semantics in the optimization process to assist the image encoder. To fully exploit shallow features, SDCL [36] introduces supervisory signals from both shallow and deep layers to guide model optimization, achieving competitive performance. Unlike the aforementioned methods, MMM [26] focuses on highlighting the importance of individual nuances in the optimization process, while mitigating the impact of noisy labels during modality-invariant learning. Although existing UVI-ReID methods have demonstrated promising performance, they overlook the impact of color variations on the acquisition of pseudo-labels. In contrast to these approaches, we propose a cross-modality augmented matching module to enhance the robustness of inter-modality pseudo-labels to color variations. Additionally, we design a soft-labels momentum update strategy, tailored to the augmented matching approach, to improve training stability.

3. Proposed Method

3.1. Overview

In the UVI-ReID task, we are given an unlabeled set of visible images $\{x_i^v\}_{i=1}^{N^v}$ and an unlabeled set of infrared images $\{x_i^r\}_{i=1}^{N^r}$, where N^v and N^r represent the number of visible and infrared images, respectively. As shown in Fig. 2, the proposed augmented and softened matching (ASM) method first performs intra-modality homogeneous learning, followed by iterations between cross-modality augmented matching (CAM), soft-labels momentum update (SMU), and inter-modality heterogeneous learning.

In intra-modality homogeneous learning, we first perform clustering on the visible and infrared image features separately to obtain intra-modality pseudo-labels. Then, we apply identity classification contrastive loss (L_{icc}^v and L_{icc}^r) to encourage the encoder to extract identity-related features within each modality. For CAM, we first apply channel augmentation (CA) [41] on the visible image set $\{x_i^v\}_{i=1}^{N^v}$ to obtain the augmented image set $\{x_i^a\}_{i=1}^{N^a}$, where $N^a = N^v$. Then, we compute centroids for all visible, augmented, and infrared clusters, followed by calculating the visible-infrared centroid similarity matrix (visible-infrared matrix) and the augmented-infrared centroid similarity matrix (augmented-infrared matrix). Finally, based on the fused similarity matrix from these two, we obtain inter-modality pseudo-labels that are robust to color variations. SMU utilizes the inter-modality pseudo-labels (e.g., $y_{i,t}^v$ and $y_{i,t}^r$) provided by CAM to update the soft pseudo-labels (e.g., $\hat{y}_{i,t}^v$ and $\hat{y}_{i,t}^r$) through momentum, enhancing training stability. In inter-modality heterogeneous learning, we introduce cross-modality soft contrastive loss (L_{csc}^v and L_{csc}^r) and cross-modality hard contrastive loss (L_{chc}^v and L_{chc}^r) to promote modality-invariant feature learning. Finally, the optimized encoder is used for the testing phase.

3.2. Intra-modality Homogeneous Learning

We first use the encoder to extract features from the visible image set $\{x_i^v\}_{i=1}^{N^v}$ and the infrared image set $\{x_i^r\}_{i=1}^{N^r}$, respectively. Then, we apply the DBSCAN algorithm [6] to cluster the visible and infrared features separately, obtaining intra-modality pseudo-labels. Subsequently, we compute the centroids of the clusters in both the visible and infrared modalities. For example, the centroid of the i-th cluster in the visible modality is defined as:

$$c_i^v = \frac{1}{N_i^v} \sum_{j=1}^{N_i^v} f_j^v, \tag{1}$$

where N_i^v represents the number of images in the i-th visible cluster, and f_j^v denotes the image feature within that cluster. Similarly, the centroid of the i-th cluster in the infrared modality is defined as:

$$c_i^r = \frac{1}{N_i^r} \sum_{j=1}^{N_i^r} f_j^r, \tag{2}$$

where N_i^r represents the number of images in the i-th infrared cluster, and f_j^r denotes the image feature within that cluster. Finally, we introduce the identity classification contrastive loss to optimize the encoder in both the visible and infrared modalities. For any image x_i^v in the visible modality, the identity classification contrastive loss is defined as:

$$L_{icc}^v = -\log \frac{\exp(f_i^v \cdot c_p^{v\mathrm{T}}/\tau)}{\sum_{q=1}^{|C^v|} \exp(f_i^v \cdot c_q^{v\mathrm{T}}/\tau)}, \tag{3}$$

where f_i^v represents the feature of x_i^v, c_p^v is the centroid of the cluster to which x_i^v belongs, $|C^v|$ is the number of clusters in the visible modality at the current epoch, and τ is the temperature hyperparameter. Similarly, for any image x_i^r in the infrared modality, the identity classification contrastive loss is defined as:

$$L_{icc}^r = -\log \frac{\exp(f_i^r \cdot c_p^{r\mathrm{T}}/\tau)}{\sum_{q=1}^{|C^r|} \exp(f_i^r \cdot c_q^{r\mathrm{T}}/\tau)}, \tag{4}$$

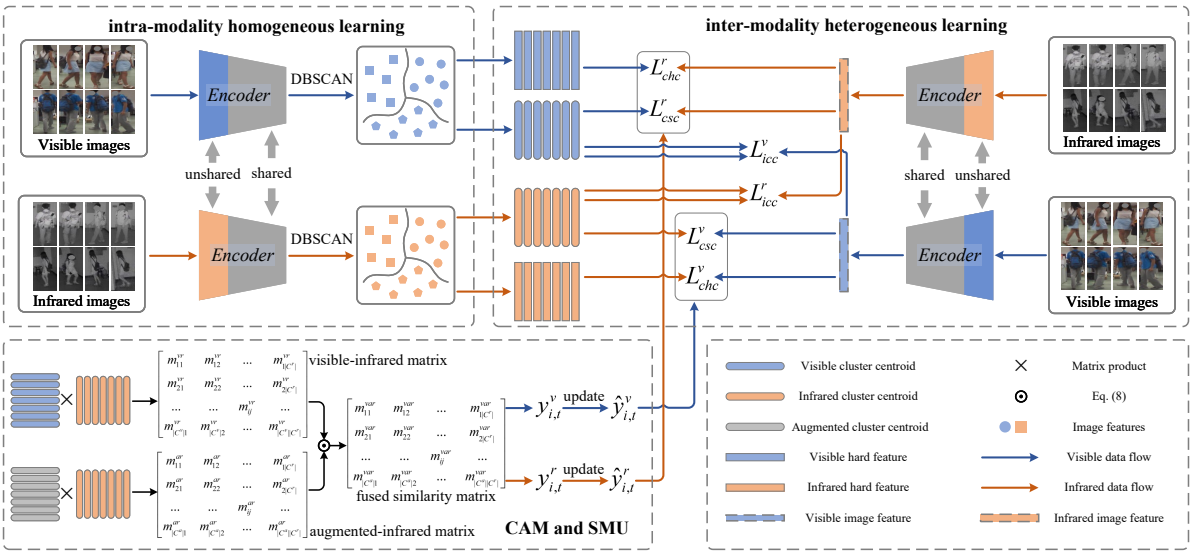


Figure 2. Illustration of the proposed augmented and softened matching (ASM) method. The encoder's frontend consists of two modality-specific branches, followed by a shared part for both modalities. Both intra-modality homogeneous learning and inter-modality heterogeneous learning include L^v_{icc} and L^r_{icc}. The different colors of the features represent different modalities, and the different shapes represent different pseudo-labels.

where f^r_i represents the feature of x^r_i, c^r_p is the centroid of the cluster to which x^r_i belongs, and $|C^r|$ is the number of clusters in the infrared modality at the current epoch.

Therefore, the total identity classification contrastive loss is defined as:

$$L_{icc} = L^v_{icc} + L^r_{icc}. \tag{5}$$

The clustering and optimization within each modality are iteratively performed. The final intra-modality pseudo-labels are saved and remain unchanged in subsequent stages.

3.3. Cross-modality Augmented Matching

To leverage data augmentation for improving the robustness of inter-modality pseudo-labels against color variations, we first apply channel augmentation (CA) [41] to the visible image set $\{x^v_i\}^{N^v}_{i=1}$, obtaining the augmented image set $\{x^a_i\}^{N^a}_{i=1}$. Then, following the strategies in Eq. (1) and Eq. (2), we calculate the centroids for all visible clusters, augmented clusters, and infrared clusters based on the intra-modality pseudo-labels saved during the homogeneous learning phase. The augmented clusters share the same pseudo-labels as the corresponding visible clusters, and the centroid of the i-th augmented cluster is denoted as c^a_i. Next, we compute the visible-infrared centroid similarity matrix $M^{vr} \in \mathbb{R}^{|C^v|\times|C^r|}$, where each element m^{vr}_{ij} in the matrix is defined as the similarity between the centroid c^v_i of the i-th visible cluster and the centroid c^r_j of the j-th infrared cluster:

$$m^{vr}_{ij} = \frac{c^v_i \cdot {c^r_j}^{\mathrm{T}}}{||c^v_i|| \cdot ||c^r_j||}. \tag{6}$$

Similarly, we compute the augmented-infrared centroid similarity matrix $M^{ar} \in \mathbb{R}^{|C^a|\times|C^r|}$, where each element m^{ar}_{ij} in the matrix is defined as the similarity between the centroid c^a_i of the i-th augmented cluster and the centroid c^r_j of the j-th infrared cluster:

$$m^{ar}_{ij} = \frac{c^a_i \cdot {c^r_j}^{\mathrm{T}}}{||c^a_i|| \cdot ||c^r_j||}. \tag{7}$$

Next, we compute the fusion matrix $M^{var} \in \mathbb{R}^{|C^a|\times|C^r|}$ by combining M^{vr} and M^{ar}. Each element m^{var}_{ij} in M^{var} is defined as:

$$m^{var}_{ij} = \frac{1}{(1+\exp(-\gamma_v \cdot m^{vr}_{ij})) \cdot (1+\exp(-\gamma_a \cdot m^{ar}_{ij}))}, \tag{8}$$

where γ_v and γ_a are shrinking factors used to control the contribution of M^{vr} and M^{ar} to M^{var}, respectively. The larger the value of $\gamma_v(\gamma_a)$, the greater the contribution of $M^{vr}(M^{ar})$ to M^{var}.

Finally, we follow the matching strategy of PGM [30] to obtain cross-modality correspondences based on the fused similarity matrix M^{var}. From the above process, it can be observed that a visible centroid is more likely to match with an infrared centroid if their similarity is high, and the corresponding augmented centroid also exhibits high similarity with the same infrared centroid. Therefore, theoretically, CAM effectively reduces the impact of color variations on matching and improves the accuracy of the matching results by leveraging the fused similarity. Relevant validations are provided in Sec. 4.4.1.

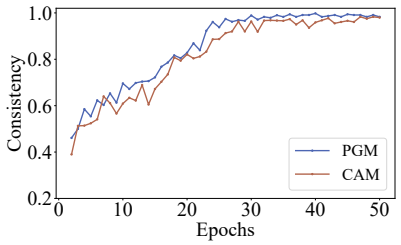


Figure 3. "Consistency" statistics on SYSU-MM01. The "Consistency" metric indicates the overlap between the matching results of the current epoch and the previous epoch.

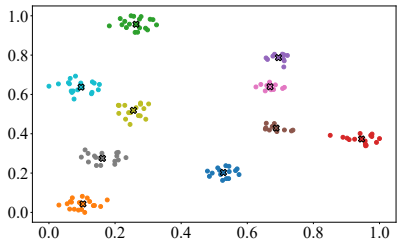


Figure 4. Visualization of intra-cluster features (dots) and centroids (crosses). Different colors represent different clusters. Some features within a cluster may be far from the centroid.

3.4. Soft-labels Momentum Update

After each CAM generates the cross-modality correspondences, each image receives a new inter-modality one-hot pseudo-label. For example, in the t-th epoch, for any visible image x_i^v, if the cluster to which it belongs matches the j-th cluster in the infrared modality, its inter-modality pseudo-label $y_{i,t}^v$ is a $|C^r|$-dimensional one-hot vector, where only the j-th element is 1, and all other elements are 0. Here, $|C^r|$ denotes the number of clusters in the infrared modality at the current epoch.

In the UVI-ReID task, cross-modality correspondences are influenced by changes in inter-cluster similarity [30], which leads to continuous variations in inter-modality pseudo-labels. As shown in Fig. 3, during the early stages of training, the visible and infrared clusters that were matched by PGM [30] in the previous epoch only have a 40% to 60% probability of being matched again in the current epoch. Moreover, the data augmentation introduced by CAM further increases the uncertainty, resulting in lower consistency during the training process. Therefore, to enhance training stability, we design a soft-labels momentum update (SMU) strategy that is adapted to CAM. Specifically, for any visible image x_i^v, we first initialize its soft-label $\hat{y}_{i,1}^v$ using its inter-modality pseudo-label $y_{i,1}^v$ from the first epoch. In each subsequent epoch, we update the soft-label using the new inter-modality pseudo-label:

$$\hat{y}_{i,t}^v \leftarrow (1-\alpha)\cdot \hat{y}_{i,t-1}^v + \alpha \cdot y_{i,t}^v, \tag{9}$$

where $\hat{y}_{i,t-1}^v$ and $\hat{y}_{i,t}^v$ represent the soft-labels of x_i^v in the $t-1$-th and t-th epochs, respectively, $y_{i,t}^v$ is the inter-modality pseudo-label of x_i^v obtained in the t-th epoch, and α is the update rate hyperparameter. For any infrared image x_i^r, the soft-label $\hat{y}_{i,t}^r$ is also updated following the same procedure as in Eq. (9).

3.5. Inter-modality Heterogeneous Learning

3.5.1. Cross-modality soft contrastive loss

To reduce the impact of modality gaps during the heterogeneous learning phase, we introduce a cross-modality soft contrastive loss. Unlike existing optimization methods [2, 30, 32, 36], the cross-modality soft contrastive loss is guided by the soft-labels provided by SMU. Specifically, for any visible image x_i^v, we first compute the similarity vector S_i^v between the image and all the cluster centroids in the infrared modality. For example, the j-th element $s_{i,j}^v$ of S_i^v is computed as the similarity between the feature f_i^v of x_i^v and the centroid c_j^r of the j-th infrared cluster:

$$s_{i,j}^v = \frac{\exp(f_i^v \cdot {c_j^r}^{\mathrm{T}}/\tau)}{\sum_{q=1}^{|C^r|}\exp(f_i^v \cdot {c_q^r}^{\mathrm{T}}/\tau)}. \tag{10}$$

Then, we compute the cross-modality soft contrastive loss as follows:

$$L_{csc}^v = -\hat{y}_i^v \cdot \log(S_i^v), \tag{11}$$

where $\hat{y}_i^v$ represents the soft-label of x_i^v. Following a similar procedure, for any infrared image x_i^r, the cross-modality soft contrastive loss is computed as:

$$L_{csc}^r = -\hat{y}_i^r \cdot \log(S_i^r), \tag{12}$$

where $\hat{y}_i^r$ and S_i^r represent the soft-label of x_i^r and the similarity vector of the feature f_i^r of x_i^r with all cluster centroids in the visible modality, respectively. During training, to further reduce the impact of noisy labels, we introduce the idea of alternating optimization [30]. Therefore, the cross-modality soft contrastive loss is defined as:

$$L_{csc} = \begin{cases} L_{csc}^v, epoch\%2 = 0 \\ L_{csc}^r, epoch\%2 = 1 \end{cases}. \tag{13}$$

3.5.2. Cross-modality hard contrastive loss

Although the cross-modality soft contrastive loss has the potential to significantly reduce modality gaps, the idea of using the cluster centroid to represent all features within a cluster overlooks the intra-cluster diversity. As shown in Fig. 4, the centroid of each cluster typically represents the shared features within the cluster, while features located at the cluster boundary often differ from the centroid. Therefore, to capture these diverse features during optimization and provide the model with more knowledge, we introduce

the cross-modality hard contrastive loss. Specifically, for each cluster in each modality, we select the image that is furthest from the cluster centroid as the hard image. Hence, each cluster has one hard image, and the number of hard images equals the number of clusters. For any visible image x_i^v, we first compute the similarity vector H_i^v between the image and all the hard images in the infrared modality. For example, the j-th element $h_{i,j}^v$ of H_i^v is calculated as the similarity between the feature f_i^v of x_i^v and the feature u_j^r of the hard image within the j-th cluster in the infrared modality:

$$h_{i,j}^v = \frac{\exp(f_i^v \cdot u_j^{r\mathrm{T}}/\tau)}{\sum_{q=1}^{|C^r|} \exp(f_i^v \cdot u_q^{r\mathrm{T}}/\tau)}. \tag{14}$$

Then, we compute the cross-modality hard contrastive loss as follows:

$$L_{chc}^v = -\hat{y}_i^v \cdot \log(H_i^v), \tag{15}$$

where $\hat{y}_i^v$ represents the soft-label of x_i^v. Following a similar procedure, for any infrared image x_i^r, the cross-modality hard contrastive loss is computed as:

$$L_{chc}^r = -\hat{y}_i^r \cdot \log(H_i^r), \tag{16}$$

where $\hat{y}_i^r$ and H_i^r represent the soft-label of x_i^r and the similarity vector between the feature f_i^r of x_i^r and the features of all hard images in the visible modality, respectively. During training, we also introduce the idea of alternating optimization [30]. Therefore, the cross-modality hard contrastive loss is defined as:

$$L_{chc} = \begin{cases} L_{chc}^v, epoch\%2 = 0 \\ L_{chc}^r, epoch\%2 = 1 \end{cases}. \tag{17}$$

3.5.3. Total loss

In summary, the total loss for inter-modality heterogeneous learning is defined as:

$$L_{total} = L_{icc} + \lambda_{csc} L_{csc} + \lambda_{chc} L_{chc}, \tag{18}$$

where λ_{csc} and λ_{chc} are the weight hyperparameters for L_{csc} and L_{chc}, respectively. The overall algorithm flow is provided in the supplementary materials.

4. Experiment

4.1. Datasets and Evaluation Metrics

We evaluate the proposed method on the SYSU-MM01 [28] and RegDB [20] datasets. SYSU-MM01 consists of 22,257 visible images and 11,909 infrared images captured by 4 visible and 2 infrared cameras. Following existing methods [30, 32, 36], we use the All Search and Indoor Search evaluation settings. In the All Search setting, the gallery contains all visible images, while in the Indoor Search setting, the gallery includes only indoor visible images. We evaluate the model on 10 galleries and report the average performance. RegDB includes 412 identities, each with 10 visible images and 10 thermal images. Following existing methods [30, 32, 36], we use the Visible to Thermal and Thermal to Visible evaluation settings. In each setting, we select all images from 206 identities for the training set and use all images from the remaining 206 identities for testing. This random selection process is repeated 10 times, and we report the average performance.

For the evaluation metrics, we report cumulative matching characteristic (CMC), mean average precision (mAP), and mean inverse negative penalty (mINP) [42].

4.2. Implementation Details

All images are resized to 288×144 pixels. During optimization, data augmentation techniques including random cropping, random flipping, random grayscale, channel random erasing [41], and channel exchange (CA) [41] are applied. For cross-modality augmented matching, only CA [41] is used for data augmentation. For the DBSCAN algorithm [6], we set the threshold and minimum number of images to 0.6 and 4, respectively, on SYSU-MM01 [28], and to 0.3 and 4 on RegDB [20]. We use AGW [42] with a ResNet-50 [12] pre-trained on ImageNet [5] as the encoder's backbone. Both intra-modality homogeneous learning and inter-modality heterogeneous learning consist of 50 epochs. Before the start of each epoch in inter-modality heterogeneous learning, we apply cross-modality augmented matching to obtain new inter-modality pseudo-labels. For each iteration, we randomly select 8 clusters from each modality, with 16 images sampled from each cluster. We use the Adam optimizer [14] to optimize the encoder. The initial learning rate is set to 0.00035, and it decays 10 times every 20 epochs. We set the temperature hyperparameter τ to 0.05. For the soft-labels momentum update, we set the update rate hyperparameter α to 0.5. For the weight hyperparameters of L_{csc} and L_{chc}, we set $\lambda_{csc} = 0.5$ and $\lambda_{chc} = 0.5$. For γ_v and γ_a, we set γ_v=2 and γ_a=1.

4.3. Comparison with State-of-the-Art Methods

As shown in Tab. 1, we first compare the proposed ASM with existing unsupervised traditional person re-identification (UT-ReID) methods, including: MMT [11], IICS [31], ICE [1], CC [4], PPLR [3], and RTMem [44]. We find that ASM significantly outperforms existing UT-ReID methods. For instance, ASM achieves 41.37% and 38.32% higher mAP and mINP, respectively, on SYSU-MM01(All Search) compared to CC [4]. This is because ASM not only introduces a cross-modality matching method but also employs optimization techniques that address modality gaps, allowing it to effectively extract modality-invariant and identity-related features.

Subsequently, we compare the proposed ASM with ex-

	Method	Reference	SYSU-MM01						RegDB					
			All Search			Indoor Search			Visible to Thermal			Thermal to Visible		
			Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP
SVI-ReID	DDAG	ECCV20	54.75	53.02	39.62	61.02	67.98	62.61	69.34	63.46	49.24	68.06	61.80	48.62
	AGW	TPAMI21	47.50	47.65	35.30	54.17	62.97	59.23	70.05	66.37	50.19	70.49	65.90	51.24
	CAJ	ICCV21	69.88	66.89	53.61	76.26	80.37	76.79	85.03	79.14	65.33	84.75	77.82	61.56
	DART	CVPR22	68.72	66.29	53.26	72.52	78.17	74.94	83.60	75.67	60.60	81.97	73.38	56.70
	SGIEL	CVPR23	77.12	72.33	-	82.07	82.95	-	92.18	86.59	-	91.07	85.23	-
	MUN	ICCV23	76.24	73.81	-	79.42	82.06	-	95.19	87.15	-	91.86	85.01	-
	SAAI	ICCV23	75.90	77.03	-	83.20	88.01	-	91.07	91.45	-	92.09	92.01	-
	IDKL	CVPR24	81.42	79.85	-	87.14	89.37	-	94.72	90.19	-	94.22	90.43	-
UT-ReID	MMT	ICLR20	**21.47**	21.53	11.50	22.79	31.50	27.66	**25.68**	**26.51**	**19.56**	**24.42**	**25.59**	**18.66**
	IICS	CVPR21	14.39	15.74	8.41	15.91	24.87	22.15	9.17	9.94	6.40	9.11	9.90	6.45
	ICE	ICCV21	20.54	20.39	10.24	**29.81**	**38.35**	**34.32**	12.98	15.64	11.91	12.18	14.82	10.60
	CC	ACCV22	20.16	**22.00**	**12.97**	23.33	34.01	30.88	11.76	13.88	9.94	11.14	12.99	8.99
	PPLR	CVPR22	11.98	12.25	4.97	12.71	20.81	17.61	10.30	11.94	8.10	10.39	11.23	7.04
	RTMem	TIP23	20.37	20.32	10.37	28.87	37.13	34.30	13.23	14.36	11.82	11.83	13.80	10.27
UVI-ReID	H2H	TIP21	30.15	29.40	-	-	-	-	23.81	18.87	-	-	-	-
	ADCA	MM22	45.51	42.73	28.29	50.60	59.11	55.17	67.20	64.05	52.67	68.48	63.81	49.62
	CHCR	TCSVT23	47.72	45.34	-	-	-	-	69.31	64.74	-	69.96	65.87	-
	CCLNet	MM23	54.03	50.19	-	56.68	65.12	-	69.94	65.53	-	70.17	66.66	-
	PGM	CVPR23	57.27	51.78	34.96	56.23	62.74	58.13	69.48	65.41	-	69.85	65.17	-
	GUR	ICCV23	63.51	61.63	47.93	71.11	76.23	72.57	73.91	70.23	58.88	75.00	69.94	56.21
	SDCL	CVPR24	64.49	63.24	51.06	**71.37**	76.90	73.50	86.91	78.92	62.83	85.76	77.25	59.57
	ASM	Ours	**65.07**	**63.37**	**51.29**	71.08	**76.91**	**73.67**	**88.23**	**79.69**	**63.07**	**86.85**	**79.20**	**59.96**

Table 1. Comparison with state-of-the-art methods on the SYSU-MM01 and RegDB datasets. The best performance under the two unsupervised settings is highlighted in bold.

isting UVI-ReID methods, including: H2H [16], ADCA [32], CHCR [21], CCLNet [2], PGM [30], GUR [35], and SDCL [36]. We find that ASM achieves a slight overall advantage over SDCL [36], which performs the best among the aforementioned methods, on SYSU-MM01, and a significant overall advantage on RegDB. This seems to be due to the more pronounced color and style differences between visible and thermal images in RegDB, where ASM's cross-modality augmented matching effectively mitigates the impact of color variations on the matching results, leading to a greater advantage over existing methods on RegDB.

Furthermore, we compare ASM with SVI-ReID methods, including DDAG [40], AGW [42], CAJ [41], DART [37], SGIEL [8], MUN [45], SAAI [7], and IDKL [24]. We find that ASM outperforms some earlier SVI-ReID methods, such as DDAG [40] and AGW [42]. Moreover, ASM's performance on RegDB is close to that of advanced SVI-ReID methods, such as SAAI [7] and IDKL [24]. This is because ASM, relying on cross-modality augmented matching, can obtain reliable cross-modality correspondences, thereby achieving results similar to those based on true labels. These experimental results validate the effectiveness and superiority of ASM.

4.4. Ablation Study

In this section, we conduct an ablation study on the important components of ASM, including cross-modality augmented matching (CAM), cross-modality soft contrastive loss L_{csc}, and cross-modality hard contrastive loss L_{chc}. Since the soft-labels momentum update (SMU) is fundamental to both L_{csc} and L_{chc}, the methods that introduce L_{csc} or L_{chc} implicitly include SMU. As shown in Tab. 2, to evaluate the effectiveness of these components, we use the existing modules PGM [30] and L_{accl} [30] as baseline components.

4.4.1. Effectiveness of CAM

As shown in Tab. 2, M2 uses CAM, rather than just PGM [30], to obtain cross-modality correspondences and achieves better performance compared to M1. To further investigate the effectiveness of CAM, we visualize some of the cross-modality correspondences for M1 and M2. As shown in Fig. 5, for visible clusters containing images with vibrant colors (e.g., red or yellow clothing), M2 is able to find more accurate cross-modality correspondences for these clusters compared to M1. This is because vibrant colors increase the appearance differences between visible images and their correct infrared counterparts. The augmented images introduced by CAM, which have a lower vibrancy compared to the visible images, are more likely to achieve higher similarity with the correct infrared images. The above experimental results validate that CAM effectively enhances the robustness of cross-modality correspondences to color variations by introducing data augmentation, thereby improving cross-modality recognition performance.

Method	PGM	CAM	L_{accl}	L_{csc}	L_{chc}	SYSU-MM01(All Search)			SYSU-MM01(Indoor Search)			RegDB(Visible to Thermal)		
						Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP
M1	✓		✓			56.87	51.62	35.03	55.53	63.13	58.88	68.51	65.03	52.89
M2		✓	✓			59.63	57.16	43.67	60.25	67.27	63.96	75.80	70.89	56.96
M3	✓			✓		58.21	53.97	38.26	57.02	63.99	60.23	71.63	67.55	55.26
M4		✓		✓		62.87	61.09	48.35	68.89	75.04	71.25	85.77	77.96	61.29
M5	✓			✓	✓	61.56	59.05	44.07	62.99	69.63	66.12	76.55	71.12	57.68
M6(ASM)		✓		✓	✓	65.07	63.37	51.29	71.08	76.91	73.67	88.23	79.69	63.07

Table 2. Ablation study on SYSU-MM01 and RegDB.

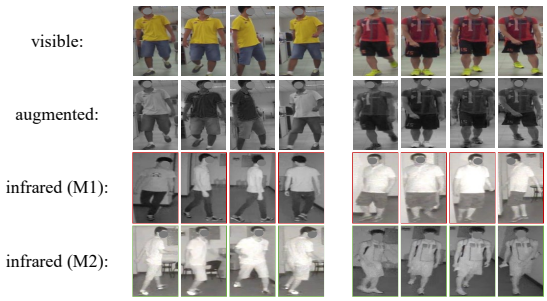
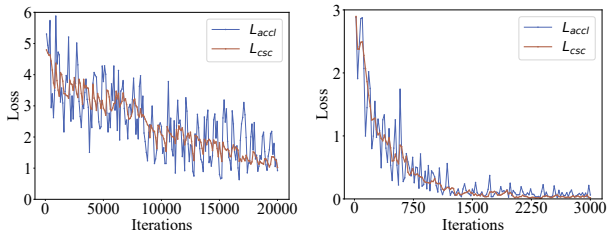


Figure 5. Visualization of cross-modality correspondences. The first four images in each row come from one cluster, and the next four images come from another cluster. The first row shows images from the visible cluster, the second row shows images from the corresponding augmented cluster, and the third and fourth rows show the infrared images matched to the visible cluster by M1 and M2, respectively. Correct and incorrect matches are marked with green and red boxes, respectively.

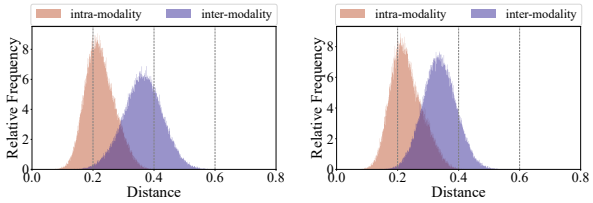


(a) SYSU-MM01 (b) RegDB

Figure 6. The loss values of L_{accl} and L_{csc} during the training process.

4.4.2. Effectiveness of L_{csc}

The cross-modality soft contrastive loss L_{csc} guides the encoder in learning modality invariance using the soft-labels provided by the SMU. As shown in Tab. 2, when L_{csc} replaces L_{accl} [30], M3 slightly outperforms M1, and M4 significantly outperforms M2. To further investigate the principles behind L_{csc}, we provide a statistical analysis of the loss values during the training process. As shown in Fig. 6, L_{csc} demonstrates higher training stability compared to L_{accl}. These experimental results confirm that L_{csc} effectively improves both training stability and model performance by utilizing soft-labels.



(a) M4 (b) M6

Figure 7. The feature distance distributions of intra-modality and inter-modality positive image pairs.

4.4.3. Effectiveness of L_{chc}

The cross-modality hard contrastive loss L_{chc} aims to leverage intra-cluster diversity to assist optimization. As shown in Tab. 2, M5 and M6 outperform M3 and M4, respectively. To further investigate the reason for the performance improvement induced by L_{chc}, we visualize the feature distances of intra-modality and inter-modality positive image pairs. As shown in Fig. 7, M6 effectively reduces the feature distance of inter-modality positive pairs compared to M4, particularly for those positive pairs with large distances, where the number of such pairs is significantly reduced. This validates that L_{chc} can effectively utilize intra-cluster diversity to further promote modality-invariant learning.

5. Conclusion

In this paper, we propose an augmented and softened matching (ASM) method for UVI-ReID. Extensive ablation experiments validate the effectiveness of the components in ASM. Specifically, cross-modality augmented matching (CAM) effectively reduces the impact of color variations on cross-modality matching by leveraging data augmentation, thereby improving the accuracy of pseudo-labels. Cross-modality soft contrastive loss utilizes the pseudo-labels provided by soft-labels momentum update (SMU) to effectively enhance training stability and model performance. Cross-modality hard contrastive loss leverages intra-cluster diversity to further promote modality-invariant learning. Comparative experimental results across four test scenarios show that ASM not only outperforms state-of-the-art unsupervised methods but also achieves competitive performance with certain supervised methods.

Acknowledgment

This work is supported by National Key Research and Development Program of China (Grant no. 2023YFC3305005).

References

[1] Hao Chen, Benoit Lagadec, and Francois Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *ICCV*, pages 14960–14969, 2021. 6

[2] Zhong Chen, Zhizhong Zhang, Xin Tan, Yanyun Qu, and Yuan Xie. Unveiling the power of clip in unsupervised visible-infrared person re-identification. In *ACM MM*, pages 3667–3675, 2023. 1, 3, 5, 7

[3] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *CVPR*, pages 7308–7318, 2022. 6

[4] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *ACCV*, pages 1142–1160, 2022. 6

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[6] 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 *KDD*, pages 226–231, 1996. 1, 3, 6

[7] Xingye Fang, Yang Yang, and Ying Fu. Visible-infrared person re-identification via semantic alignment and affinity inference. In *ICCV*, pages 11270–11279, 2023. 2, 7

[8] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *CVPR*, pages 22752–22761, 2023. 7

[9] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE TIP*, 29:579–590, 2019. 2

[10] Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, and Ran He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In *ICCV*, pages 11823–11832, 2021. 1

[11] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *ICLR*, 2020. 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[13] Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In *AAAI*, pages 1034–1042, 2022. 1

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 6

[15] Xulin Li, Yan Lu, Bin Liu, Yating Liu, Guojun Yin, Qi Chu, Jinyang Huang, Feng Zhu, Rui Zhao, and Nenghai Yu. Counterfactual intervention feature transfer for visible-infrared person re-identification. In *ECCV*, pages 381–398. Springer, 2022. 1

[16] Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. *IEEE TIP*, 30:6392–6407, 2021. 2, 7

[17] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *CVPR*, pages 19366–19375, 2022. 1

[18] Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *AAAI*, pages 1835–1843, 2023. 1

[19] 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 *CVPR*, pages 13379–13389, 2020. 2

[20] 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. 6

[21] 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 TCSVT*, 2023. 3, 7

[22] Zhihao Qian, Yutian Lin, and Bo Du. Visible-infrared person re-identification via patch-mixed cross-modality learning. *PR*, 157:110873, 2025. 2

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2, 3

[24] Kaijie Ren and Lei Zhang. Implicit discriminative knowledge learning for visible-infrared person re-identification. In *CVPR*, pages 393–402, 2024. 7

[25] Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *ICCV*, pages 11218–11228, 2023. 1

[26] 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 *ECCV*, pages 456–474. Springer, 2024. 3

[27] Jiangming Shi, Xiangbo Yin, Yachao Zhang, Yuan Xie, Yanyun Qu, et al. Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification. In *NeurIPS*, pages 99715–99734, 2024. 1

[28] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. RGB-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017. 2, 6

[29] Jianbing Wu, Hong Liu, Yuxin Su, Wei Shi, and Hao Tang. Learning concordant attention via target-aware alignment

for visible-infrared person re-identification. In *ICCV*, pages 11122–11131, 2023. 1

[30] Zesen Wu and Mang Ye. Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In *CVPR*, pages 9548–9558, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[31] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *CVPR*, pages 11926–11935, 2021. 6

[32] Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *ACM MM*, pages 2843–2851, 2022. 1, 2, 3, 5, 6, 7

[33] Bin Yang, Jun Chen, Cuiqun Chen, and Mang Ye. Dual consistency-constrained learning for unsupervised visible-infrared person re-identification. *IEEE TIFS*, 2023.

[34] Bin Yang, Jun Chen, Xianzheng Ma, and Mang Ye. Translation, association and augmentation: Learning cross-modality re-identification from single-modality annotation. *IEEE TIP*, 2023.

[35] Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *ICCV*, pages 11069–11079, 2023. 2, 7

[36] Bin Yang, Jun Chen, and Mang Ye. Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In *CVPR*, pages 16870–16879, 2024. 1, 2, 3, 5, 6, 7

[37] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *CVPR*, pages 14308–14317, 2022. 7

[38] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, page 2, 2018. 1, 2

[39] Mang Ye, Xiangyuan Lan, and Qingming Leng. Modality-aware collaborative learning for visible thermal person re-identification. In *ACM MM*, pages 347–355, 2019. 2

[40] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *ECCV*, pages 229–247. Springer, 2020. 1, 2, 7

[41] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *ICCV*, pages 13567–13576, 2021. 1, 2, 3, 4, 6, 7

[42] 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 TPAMI*, 44(6): 2872–2893, 2021. 1, 6, 7

[43] Mang Ye, Zesen Wu, Cuiqun Chen, and Bo Du. Channel augmentation for visible-infrared re-identification. *IEEE TPAMI*, 2023. 1

[44] Junhui Yin, Xinyu Zhang, Zhanyu Ma, Jun Guo, and Yifan Liu. A real-time memory updating strategy for unsupervised person re-identification. *IEEE TIP*, 32:2309–2321, 2023. 6

[45] Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *ICCV*, pages 11185–11195, 2023. 1, 7

[46] Xiaoyan Yu, Neng Dong, Liehuang Zhu, Hao Peng, and Dapeng Tao. Clip-driven semantic discovery network for visible-infrared person re-identification. *IEEE TMM*, 2025. 2

[47] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. FMCNet: Feature-level modality compensation for visible-infrared person re-identification. In *CVPR*, pages 7349–7358, 2022. 2

[48] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 3

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 2, 3