Cross-Modality Person Re-Identification via Modality Confusion and Center Aggregation

Xin Hao[1], Sanyuan Zhao[1], Mang Ye[2]*, Jianbing Shen[3]
[1] School of Computer Science, Beijing Institute of Technology, Beijing, China
[2] School of Computer Science, Wuhan University, Wuhan, China
[3] State Key Laboratory of IoT for Smart City, Department of Computer and Information Science, University of Macau, Macau, China
haoxin@bit.edu.cn, yemang@whu.edu.cn

Abstract

Cross-modality person re-identification is a challenging task due to large cross-modality discrepancy and intra-modality variations. Currently, most existing methods focus on learning modality-specific or modality-shareable features by using the identity supervision or modality label. Different from existing methods, this paper presents a novel Modality Confusion Learning Network (MCLNet). Its basic idea is to confuse two modalities, ensuring that the optimization is explicitly concentrated on the modality-irrelevant perspective. Specifically, MCLNet is designed to learn modality-invariant features by simultaneously minimizing inter-modality discrepancy while maximizing cross-modality similarity among instances in a single framework. Furthermore, an identity-aware marginal center aggregation strategy is introduced to extract the centralization features, while keeping diversity with a marginal constraint. Finally, we design a camera-aware learning scheme to enrich the discriminability. Extensive experiments on SYSU-MM01 and RegDB datasets show that MCLNet outperforms the state-of-the-art by a large margin. On the large-scale SYSU-MM01 dataset, our model can achieve 65.40 % and 61.98 % in terms of Rank-1 accuracy and mAP value.

1. Introduction

Person re-identification (ReID) is a technique that uses computer vision technology to determine whether there is a specific person from a gallery set captured by surveillance cameras [17]. It has gained increasing attention in computer vision area for both research and application. However, there are relatively few works paying attention to the ReID between visible images and infrared images.

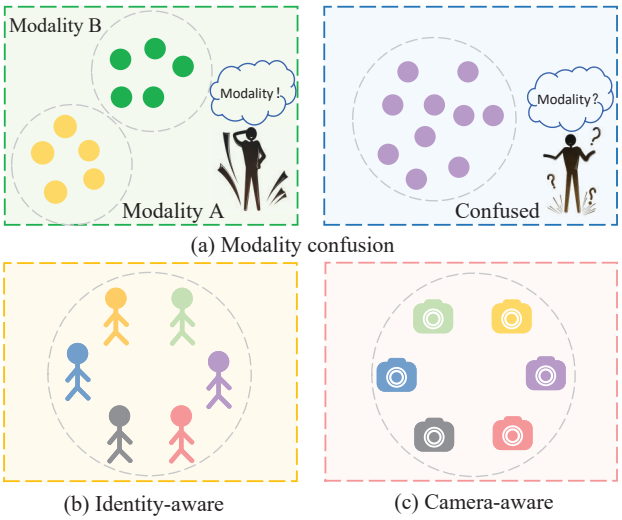


Figure 1. Idea illustration. (a) is the modality confusion learning process. After that, the two modalities are difficult to be correctly classified, narrowing the modality discrepancy. (b) and (c) are the designed identity-aware and camera-aware marginal constrained center aggregation for person ID and camera ID prediction.

This cross-modality visible-infrared person re-identification (cm-ReID) [40] problem is also an important issue in night-time surveillance application. Compared to the widely studied single-modality ReID [5, 52], the cm-ReID is much more challenging due to large visual differences between the two modalities and different camera environments.

To narrow the gap between two modalities, existing methods mainly focus on learning shareable common feature representations, via either one [37, 40] or two-stream networks [45, 9]. [15] designs a spectrum dispelling branch to eliminate the influence of the spectrum. Besides, some methods generate a common intermediate modality [16] to eliminate the influences caused by modality discrepancy. A similar approach adopts GAN technique [10] to generate cross-modality images for person matching. However, gen-

*Corresponding Author: *Mang Ye*

erating common modality or generating cross-modality images is unavoidably accompanied by noises. Worse still, the generated images greatly increase the computational burden and add more uncertainty for the cross-modality learning, limiting the applicability for practical model deployment.

To address the above limitations, we propose a novel end-to-end Modality Confusion Learning network (MCLNet), which aims to learn modality-invariant features. Our network neither needs the prior information of input nor generates additional subspace features, which ensures the maximization of input information without additional noise. The basic idea is to confuse the modality discrimination in the feature learning process, making the optimization explicitly focusing on modality-irrelevant perspective (Fig. 1 (a)). MCLNet minimizes inter-modality discrepancy while maximizing cross-modality similarity among instances through the min-max game [27]. Incorporated with a partially shared two stream network, our MCLNet can simultaneously learn the modality-specific features and extract the modality-invariant features. In a confusion learning manner, it achieves a balance between the modality confusion and the general cross-modality feature learning.

Furthermore, we introduce an identity-aware marginal center aggregation strategy to reinforce the representation invariance against modality discrepancy (Fig. 1 (b)). The basic idea is to constrain that samples belonging to the same identity across two modalities are invariant. While encouraging the extraction of centralized features, a marginal constraint is incorporated to make sure that the samples are not too concentrated. This keeps the feature diverse, but it greatly enhances the generalization ability. In addition, based on the observation that person images are captured in totally different camera environments, we further integrate a camera-aware marginal center aggregation scheme (Fig. 1 (c)). This component fully utilizes the camera labels, capturing the camera-specific information for the learned representation. This constraint enhances the robustness against camera variations. The proposed components might be easily integrated into other advanced learning models.

Our main contributions can be summarized as follows: We propose a novel Modality Confusion Learning Network (MCLNet) for cm-ReID. It is an effective learning structure to extract modality irrelevant representation, reinforcing the robustness of the learned representation against modality variations. We introduce an identity-aware marginal constrained center aggregation strategy. It extracts the centralization features, while keeping the diversity for better generalization ability with a marginal constraint. We design a camera-aware learning scheme that applies the camera label supervision, enriching the discriminability via camera-aware representations. Extensive experimental results show that our novel framework outperforms the state-of-the-art methods on two cm-ReID datasets.

2. Related Work

2.1. Single-Modality Person Re-Identification

Single-modality person re-identification aims at matching the person images captured by different cameras in the daytime, while all the images are from the same visible modality. Existing works have shown desirable performance on the widely-used datasets with deep learning technique [50, 18, 58, 14, 32, 2, 59, 20]. A few methods propose to solve person re-identification as ranking problems [25, 3]. Some methods are aware of local information and global information [35, 31, 30, 53], which can improve the performance when they are combined. Besides, [59, 5, 36, 42, 53] focused on the loss functions designed for deep learning. [51, 4, 54, 41] utilize attention information to enhance representation learning. Some methods attempt to solve person re-identification problems using domain adaptation methods [57, 8], since images from each camera can be regarded as an independent domain. However, in practical applications, most cameras switch modes between visible and infrared during the day and night. Due to the large cross-modality discrepancy, single modality solutions are no longer competent for cm-ReID task, leading to poor generalization performance.

2.2. Cross-Modality Person Re-Identification

To reduce cross-modality discrepancy, [40] proposes a deep zero-padding network to extract useful embedding features. Two-stream networks [45, 47, 44, 46] can learn both the modality-shared features and the modality-specific information. [26] applies a dual Gaussian-based variational auto-encoder, to disentangle an identity-discriminable and an identity-ambiguous cross-modality feature subspace. [19] proposes a cross-modality shared-specific feature transfer algorithm to explore the potential of both the modality-shared information and the modality-specific characteristics. [13] exploits the intra-modality sample similarities to circumvent the cross-modality image matching. A modality-aware learning method [43] handles the modality discrepancy at the classifier level. [16] designs an auxiliary X-modality to mitigate the influence of modality discrepancy. Generative adversarial networks have been adopted in cm-ReID, by generating data to mitigate the modality discrepancy. [37] generates cross-modality images from two different modality images, and combines the generated images and the real images to bring about mixed multi-spectral images. An end-to-end alignment generative adversarial network [33] exploits pixel alignment and feature alignment jointly. [7] applies GAN to handle the lack of insufficient discriminative information and the issue of large scale cross-modality metric learning. [34] generates cross-modality paired-images and performs both global set-level and fine-grained instance-level alignments.

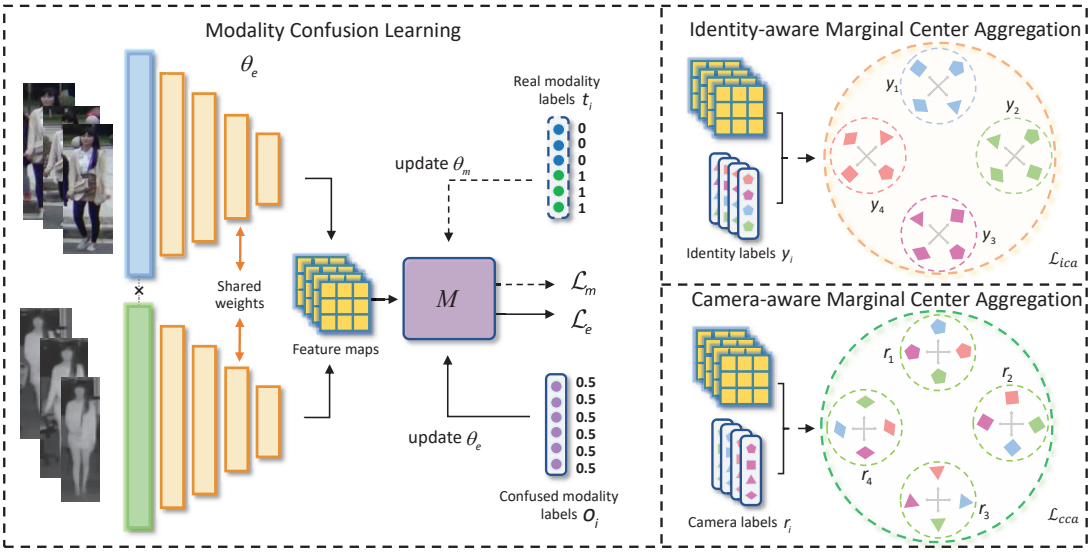


Figure 2. Illustration of the modality confusion Learning network (MCLNet). The cross-modality images are fed into a feature extractor, which confuses the modality feedback through a confusion learning mechanism. We also incorporate an identity-aware and a camera-aware marginal center aggregation strategy to further enhance the discriminability by using both the identity and camera labels (labels with the same color mean they have the same person ID, and those with the same shape mean they have the same camera ID).

Differently, we propose a modality confusion learning network, which can learn modality-invariant features by minimizing inter-modality discrepancy while maximizing cross-modality similarity among instances. Compared with above methods, it does not need to generate cross-modality images, and simultaneously takes into account the modality robustness features. In addition, our method uses only global features to achieve the best performance.

3. Proposed Method

The Modality Confusion Learning Network (MCLNet) consists of three major components as shown in Fig. 2. It is designed on top of a partially shared two-stream network for modality-invariant feature learning (§ 3.1). First, modality confusion learning module confuses the modality discrimination feedback for better modality-irrelevant property (§ 3.2). Then, we present an identity-aware marginal center aggregation strategy (§ 3.3) to improve the identity-centralized representation learning. Finally, camera-aware marginal center aggregation constraint (§ 3.4) is presented by exploiting the camera label information to learn camera-aware representations.

3.1. Feature Extractor

We adopt a generic framework of person ReID, named AGW [47], as our baseline. Our feature extractor is a two-stream network successively extracting modality-specific and modality-shared features. Specifically, to handle the discrepancy of two heterogeneous modalities, in the first convolutional block, the visible and the infrared images are processed independently, aiming at learning low-level features. After that, the following four blocks of the two streams share parameters and extract high-dimensional features in common. This architecture benefits from a uniform structure that simultaneously captures the cross-modality information and generates common used features that can be processed in our single framework.

3.2. Modality Confusion Learning

From the baseline, the extracted features may depend on the modalities, such as color or spectrum. Therefore, the visible and the infrared samples have different feature distributions and will not be well aligned for comparison. To reduce the discrepancy between the visible and infrared images, our network is designed to ignore the information of the modality and learn common representations for person. However, "common" is not equal to "useful". If we merely mystify the network about the sample modality during training, the network may focus on trivial features and overlook the particular features of different persons, which leads to the failure of useful information collection.

Considering the above two points, we design a confusion learning mechanism, the inter-modality discrepancy is minimized and the cross-modality similarity is maximized via the min-max game. Thus, while learning the modality-irrelevant features, the network is constrained to pick up discriminative component to predict person identities. Different from existing methods utilizing GAN [7, 37, 33, 34] to transfer sample domains between modalities by generating cross-modality images, we apply the confusion learn-

ing mechanism for deceiving the network into confounding the visible modality with the infrared one. This mechanism avoids the risk of poor quality and noises of the generated cross-modality images, and operates on the embeddings of the two modalities directly. Specifically, our goal is to achieve a confusion that the modality classifier cannot distinguish the modality of an input image.

Formally speaking, for each sample image x_i, there is an identity label y_i, a real modality label t_i and a confused modality label o_i. Specifically, we use a two-dimensional vector to define the one-hot modality label. For each input sample x_i, the real modality label t_i is set to [1, 0] for visible image and [0, 1] for the infrared image. For the confused modality label o_i, it is set to [0.5, 0.5] for all the samples from two different modalities. Our modality confusion learning requires two components: feature extractor and modality confusion module M. We represent M with parameters θ_m to act as **M**odality **C**onfusion **M**odule (MCM). It is essentially a two-layer classifier and its purpose is to accurately distinguish the input images into a certain modality. For sample x_i with extracted feature f_{x_i}, M outputs the modality prediction probability $p_m(f_{x_i})$, and we compare it with the real modality label t_i. The loss function of M can be formulated as:

$$\mathcal{L}_m(\theta_m) = -\frac{1}{N}\sum_{i=1}^{N} t_i \cdot \log p_m(f_{x_i}, \theta_m; \theta_e), \quad (1)$$

where N denotes the sample number in a batch, x_i is the i^{th} input sample. Given a learned feature extractor θ_e and the modality classifier θ_m, the probability of sample x_i being correctly classified is represented by $p_m(f_{x_i}, \theta_m; \theta_e)$, normalized by a softmax function.

The purpose of feature extractor is to extract features that are modality-invariant and discriminative. Similarly, we construct E with parameters θ_e to act as feature extractor. To achieve modality confusion, we compare the predicted probability of the feature extractor with the confused modality label o_i. And the loss function can be formulated as:

$$\mathcal{L}_e(\theta_e) = -\frac{1}{N}\sum_{i=1}^{N} o_i \cdot \log p_m(f_{x_i}, \theta_e; \theta_m), \quad (2)$$

In the training stage, we update θ_m and θ_e alternately until they reach equilibrium. θ_e represents the feature extractor, which aims to maximize the loss of the modality confusion module by making the feature distributions as similar as possible. θ_m means the modality confusion module, which aims to minimize the loss of the modality classifier to help the network distinguish modality. θ_m and θ_e can be optimized as follows:

$$\begin{aligned} \mathcal{L}(\theta_m, \theta_e) &= \mathcal{L}_m(\theta_m) + \mathcal{L}_e(\theta_e) \\ \hat{\theta}_m &= \arg\min_{\theta_m} \mathcal{L}(\theta_m, \hat{\theta}_e) \\ \hat{\theta}_e &= \arg\min_{\theta_e} \mathcal{L}(\hat{\theta}_m, \theta_e). \end{aligned} \quad (3)$$

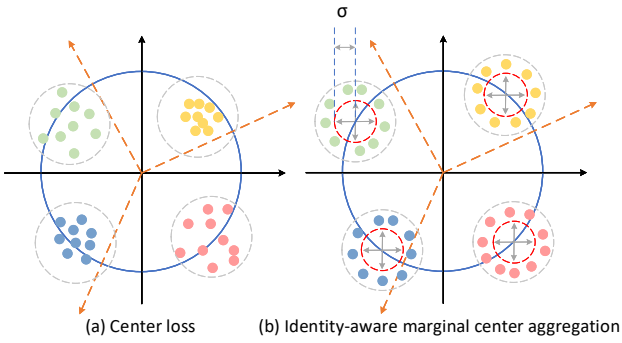


Figure 3. Comparison between (a) center loss and (b) identity-aware marginal center aggregation. Different colors represent embedding features from different identities. σ is the predefined hyperparameter for the margin.

In the optimization process, one module will be updated at each step while the other component will be fixed [18]. This strategy will ensure that the network updates in the correct gradient. Our target is that the embeddings extracted by feature extractor can not be correctly classified into the corresponding modality, achieving the modality confusion.

3.3. Identity-aware Marginal Center Aggregation

Similar to single-modality person re-identification, the appearance of persons in cm-ReID is also easily affected by clothing, scale, shielding, attitude and viewpoints [49, 29], which makes the ReID task more difficult. To handle this problem, most existing methods adopt center loss [21] to simultaneously learn a center of each class for feature embeddings, and penalize the distances between the samples and their corresponding classes. Center loss [38] can be represented as:

$$\mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{N} \|f_{x_i} - c_{y_i}\|_2^2, \quad (4)$$

where x_i denotes the i^{th} input sample, c_{y_i} is the ${y_i}^{th}$ class center, f_{x_i} is the embedding extracted by feature extractor.

Center loss was firstly applied to solve the face recognition problem [38] and achieved good performance. The main reason is that faces have strong centrality. However, person characteristics are influenced by many factors, especially when cm-ReID suffers from large cross-modality gap. Strictly concentrating the images of the same identity from two modalities will sacrifice the diversity of varying person images, leading to limited generalization ability on the testing set. Considering it, we propose an **I**dentity-aware marginal **C**enter **A**ggregation strategy (ICA) to extract the centralized features temperately and take the discriminative information into account.

As demonstrated in Fig. 3, each color corresponds to a certain identity. Center loss draws all the samples close to the center of corresponding class (Fig. 3 (a)). Differently,

ICA constrains features of the same identity within a certain range (Fig. 3 (b)). A constraint margin σ is applied to ensure the samples belonging to the same class not too close to the center, preventing the features from overfitting to the feature center, to moderately keep the diversity of the identity description, even in different modalities. This diversity can provide more sample-specific information for the network to distinguish different person identities. ICA encourages features of the same identity distributed on a high-dimensional sphere uniformly, rather than blindly pursuing representation similarity. The loss of ICA for identity prediction can be represented as:

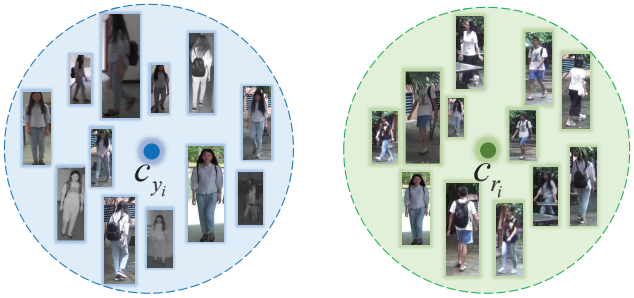
$$\mathcal{L}_{ica} = \frac{1}{N}\sum_{i=1}^{N}[||f_{x_i} - c_{y_i}||_2^2 - ||f_{y_i}^o - c_{y_i}||_2^2 - \sigma]_+ + \frac{1}{N}\sum_{i=1}^{N}[\rho - \min_{y_j \neq y_i} ||c_{y_i} - c_{y_j}||_2^2]_+ \tag{5}$$

where x_i denotes the i^{th} sample, y_i is the identity label of x_i, c_{y_i} is the class center of ${y_i}^{th}$ identity, $f_{y_i}^o$ is the feature embedding which is the closest to center c_{y_i}, N is the sample number in a batch, σ is the hyper-parameter denoting the radius of the sphere of a certain class. In Eq. 5, on the right side of the equation, the first element in the first term is a general form of center loss, which indicates the outer circle (in gray color) constraint in Fig. 3 (a) and (b). The second element is the minimum distance between samples and center c_{y_i}. It is subtracted in order to push the sample features gradually away from center c_{y_i} by a small margin. The hyper-parameter σ forces a sample to keep a reasonable distance from its identity center. σ can be viewed as the radius of the inner circle (in red color) in Fig. 3 (b). This design avoids too strict center concentration. The second term calculates the minimum distance of a different identity center. By applying multiple constraints between different identity centers, the network compares identity similarity rather than sample similarity.

3.4. Camera-aware Marginal Center Aggregation

Considering the large camera difference, this section presents a strategy to exploit the camera label information for further improvement, reinforcing the modality-invariant features learning.

In real life, cm-ReID tasks are usually captured by multiple cameras. This motivates us to model camera differences for the following reasons: 1) Different camera internal parameters are different. 2) Different cameras have different backgrounds and viewing angles. 3) There is usually no overlapping area between cameras. As a result, we propose a **C**amera-aware marginal **C**enter **A**ggregation strategy (CCA). Our goal is to let the network learn discriminative information about different cameras. Specifically, we expect the network to also pay attention to the differences in images from different cameras, due that these cameras usually work in different modes or in different environments. With the common constraint of ICA and CCA, the network is encouraged to mine the implicit identity association information between the same person under different cameras. The camera-aware marginal center aggregation loss can be represented as:



(a) Identity center (b) Camera center

Figure 4. Comparison pictures of identity-aware marginal center aggregation and camera-aware marginal center aggregation. (a) is the images from the different cameras within the same identity. (b) is the images from the different identities within the same camera. c_{y_i} is the ${y_i}^{th}$ identity center, c_{r_i} is the ${r_i}^{th}$ camera center.

$$\mathcal{L}_{cca} = \frac{1}{N}\sum_{i=1}^{N}[||f_{x_i} - c_{r_i}||_2^2 - ||f_{r_i}^o - c_{r_i}||_2^2 - \sigma]_+ + \frac{1}{N}\sum_{i=1}^{N}[\rho - \min_{r_j \neq r_i} ||c_{r_i} - c_{r_j}||_2^2]_+, \tag{6}$$

where r_i denotes the camera label of i^{th} sample and c_{r_i} is the ${r_i}^{th}$ camera center. $f_{r_i}^o$ represents the closest sample to the camera center. Other elements are similar to Eq. 5.

Identity-aware and camera-aware marginal center aggregation strategy extract discriminative information in different ways. As shown in Fig. 4, on the one hand, ICA constrains the same identity image from different cameras. On the other hand, CCA constrains the same camera image from different identities. These two components work together to explicitly apply identity-specific information and camera-specific information.

Overall. Certainly, a conventional loss function consisting of identity loss (ID loss [55]) $\mathcal{L}_{id}$ and weighted regularization triplet (WRT) loss [47] $\mathcal{L}_{wrt}$ is designed to learn dicriminative representation:

$$\mathcal{L}_b = \mathcal{L}_{id} + \mathcal{L}_{wrt}. \tag{7}$$

Identity loss utilizes cosine distance to separate the embeded space into different subspaces. Triplet loss enhances the intra-class compactness and inter-class separability in the Euclidean space. Distribution in the embedding space is supervised by ID loss and triplet loss in different dimensions, so that the model could learn more discriminative features. After feature extraction, feature embeddings achieving desirable performance are obtained. We have adopted ICA

and CCA for person identity prediction and camera identity prediction, respectively. In summary, the final loss is:

$$\mathcal{L}_{total} = \mathcal{L}_b + \mathcal{L}_e + \lambda(\mathcal{L}_{ica} + \mathcal{L}_{cca}), \tag{8}$$

where λ is a predefined trade-off parameters. It is worth noting that $\mathcal{L}_m$ is not included in total loss since the updates of $\mathcal{L}_e$ and $\mathcal{L}_m$ are separated. They are updated alternately through adversarial training, supervising the network to achieve modality confusion. When the modality confusion is achieved, we can ignore $\mathcal{L}_m$.

4. Experiment and Analysis

In this section, we evaluated our model on two public cm-ReID datasets, SYSU-MM01 [40] and RegDB [24].

4.1. Datasets and Evaluation Protocol

SYSU-MM01 [40] is the first large-scale benchmark dataset for cm-ReID collected by 6 cameras, consisting of 4 visible and 2 infrared cameras. Specially, four cameras are placed in the outdoor environments and two are in the indoor environments. SYSU-MM01 contains 491 persons with a total of 287,628 visible images and 15,792 infrared images. 395 persons including 22,258 visible images and 11,909 infrared images for training, another 96 persons including 3803 infrared images for query and 301 randomly selected visible images as gallery set. Meanwhile, it contains two different testing modes, *all-search* and *indoor-search* modes. Detailed description of the experimental settings can be found in [40].

RegDB [24] is collected by a dual-camera system, including one visible and one infrared camera. This dataset includes 412 persons, for each of them, there are 10 visible images and 10 infrared images. Following the evaluation protocol proposed by [44], we randomly select 206 identities for training and another 206 identities for testing. The testing stage contains two modes, *visible-to-infrared* and *infrared-to-visible*, which means that the images from one modality were used as the gallery set while the remaining as the query set. The results are averaged for 10 trials to obtain stable performance [37].

Evaluation Protocol. The cumulative matching characteristics (CMC) [23], mean average precision (mAP) and mean inverse negative penalty [47] (mINP) are used as evaluation metrics.

4.2. Implementation Details

This work is supported by Huawei MindSpore [1]. MCLNet adopts AGW [47] as feature extractor. Before the training stage, batchsize is set to 64, containing 32 visible and 32 infrared images from 8 identities. For each identity, 4 visible and 4 infrared images are selected randomly. Both modalities images are from the original three channels. The input images are first resized to 288 × 144, then we adopt random crop with zero-padding, random horizontal flipping and random erasing for data augmentation [56]. The finally cropped image sizes are 256 × 128 for both modalities. In the training stage, MCM updates one time while feature extractor updates once. In addition, we use the SGD optimizer for optimization, and the momentum parameter is set to 0.9. The warm-up learning rate is adopted which is initially set to 0.1. We decay it by 0.1 and 0.01 at 20 and 50 epochs. The whole training process consists of 200 epochs. We set the predefined parameter $\lambda = 0.0005$ for $\mathcal{L}_{ica}$ and $\mathcal{L}_{cca}$, λ is used to balance the contributions of different losses due to its large value.

Table 1. Analysis of the effectiveness of modality confusion learning on SYSU-MM01 dataset under the *all-search* mode. Rank-1 accuracy (%), mAP (%) and mINP (%) are reported. AGW* means AGW uses random erasing [56].

Methods	Rank-1	mAP	mINP
Base	49.40	49.02	35.82
Base+MCM	51.46	49.84	36.73
AGW	47.50	47.65	35.30
AGW+MCM	49.29	49.26	37.08
AGW*	59.82	56.07	40.50
AGW*+MCM	**62.74**	**58.83**	**43.15**
MCLNet	**65.40**	**61.98**	**47.39**

4.3. Ablation Study

In this subsection, we evaluate the effectiveness of each component of our proposed method.

Effectiveness of Modality Confusion Module. Firstly, we evaluate how much improvement can be made by modality confusion mechanism on the SYSU-MM01 dataset under the *all-search* mode. It is worth noting that our feature extractor could be replaced with most existing cm-ReID embedding features extraction networks. We study this nice property by applying modality confusion mechanism to Base and AGW [47]. Base uses ResNet-50 [12] pre-trained on ImageNet [28] as backbone. For a fair comparison, we change ResNet-50 to have the same two-stream structure as AGW. Meanwhile, they use only $\mathcal{L}_{id}$ and $\mathcal{L}_{wrt}$ in the training stage. As Table 1 shows, the performance results are improved when both feature extractor and MCM are incorporated. Meanwhile, our modality confusion mechanism is still effective after using data augmentation. AGW* + MCM achieves a rank-1 accuracy of 62.74%, a mAP of 58.83% and a mINP of 43.15% which are higher than AGW* by 2.92%, 2.76% and 2.65%, respectively.

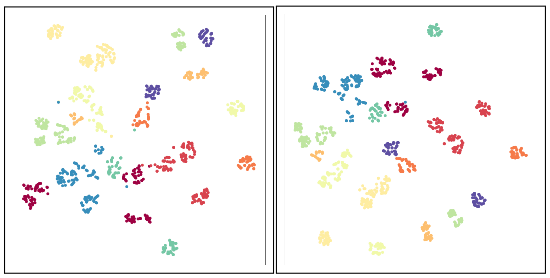
Effectiveness of ICA and CCA. Secondly, we performed a comparative experiment on the ICA and CCA on the SYSU-MM01 dataset (*all-search* mode) to verify its validity. As the Table 2 shows, MCLNet with ICA and CCA achieve 4.31%, 4.60% and 5.56% improvements in Rank-1, mAP and mINP, respectively. When the ICA and CCA are respectively used for person ID prediction and camera

Table 2. Analysis of the effectiveness of ICA and CCA on SYSU-MM01 dataset under the *all-search* mode. Rank-1 accuracy(%), mAP(%) and mINP(%) are reported.

Methods	Rank-1	mAP	mINP
AGW*	59.82	56.07	40.50
AGW* + ICA	63.56	59.77	44.45
AGW* + CCA	63.42	59.19	44.13
AGW* + ICA, CCA	64.13	60.67	46.06
MCLNet	**65.40**	**61.98**	**47.39**

Table 3. Effectiveness of ICA and CCA over different baselines on SYSU-MM01 dataset under the *all-search* mode. Rank-1 accuracy(%), mAP(%) and mINP(%) are reported.

Methods	Rank-1	mAP	mINP
Base	49.40	49.02	35.82
Base + ICA, CCA	52.13	50.89	37.96
DDAG [46]	54.75	53.02	39.62
DDAG + ICA, CCA	57.27	54.32	40.03
AGW* [47]	59.82	56.07	40.50
AGW* + ICA, CCA	64.13	60.67	46.06



(a) AGW* (b) MCLNet

Figure 5. The t-SNE [22] visualization of features on SYSU-MM01 dataset. We randomly select 10 identities of testing set and extract person descriptors use AGW* and MCLNet, respectively. Different colors represent different identities.

ID prediction, the results are improved, and the results are better when they are combined. It can be observed that ICA and CCA have more improvements on mAP and mINP, which indicates it is easier for the framework to find the images of the same identity, validating that ICA and CCA are suitable for cm-ReID task.

Visualization Analysis. To further analyze the effectiveness of MCLNet, we use t-SNE [22] to transform high-dimensional features vectors into two-dimensional vectors. As shown in Fig. 5, compared to the visualization results of AGW*, the features extracted from MCLNet are better clustered together. The distance between the centers and boundaries among different identities are more obvious, verifying that our work is more discriminating.

4.4. Parameters Analysis

The proposed MCLNet involves two key parameters, including ICA/CCA balanced weight λ and ICA/CCA margin σ. The two parameters are studied by setting them to different values as shown in Fig. 6 and Fig. 7, respectively. On the one hand, due to the large value of $\mathcal{L}_{ica}$ and $\mathcal{L}_{cca}$, the value of λ is set to match well with $\mathcal{L}_{id}$ and $\mathcal{L}_{tri}$ to

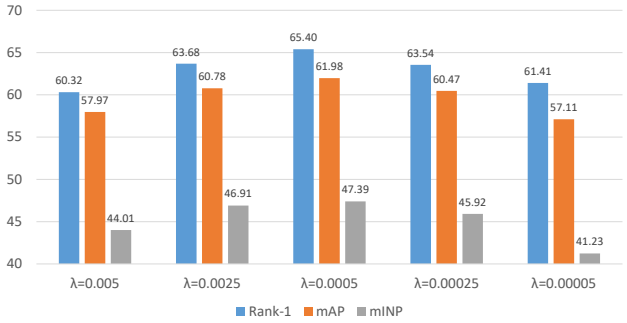


Figure 6. The effect of parameter λ on SYSU-MM01 dataset under the *all-search* mode. λ is used to balance the contributions of different losses due to its large value.

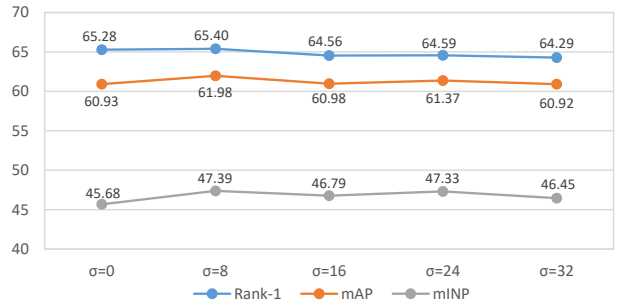


Figure 7. The effect of parameter σ on SYSU-MM01 dataset under the *all-search* mode. Rank-1, mAP and mINP (%) are reported.

balance the contributions and ensure that they converge together. On the other hand, the value of σ indicates how far the embeddings need to be from the centers. It ensures the samples belonging to the same class not too close to the center. We can observe that the introduction of σ has improved mAP and mINP, verifying our conjecture that σ can maintain feature diversity, and make the network stronger ability to retrieve the hardest correct match. However, σ should be a reasonable margin to prevent the sacrifice of Rank-1 accuracy. Experiments show that MCLNet performs optimally when $\lambda = 0.0005$, $\sigma = 8$.

4.5. Comparison With the State-of-the-art Methods

In this section, the proposed MCLNet is compared with state-of-the-arts on two different datasets, including more than ten competing methods publsihed in recent two years. The results are listed in Table 4 and 5, respectively.

The experiments on SYSU-MM01 dataset (Table 4) show that MCLNet achieves competitive performance compared with the state-of-the-arts. According to the experimental results, the following observations can be made: 1) Our method performs much better than the methods (cmGAN [7], AliGAN [33], XIV [16], Hi-CMD [6]) that generate cross-modality image pairs by GAN or utilize auxiliary modality. Meanwhile, MCLNet does not require time-

Table 4. Comparison with the state-of-the-arts on SYSU-MM01 dataset. Rank-k accuracy (%), mAP (%) and mINP (%) are reported.

Settings		*All Search*					*Indoor Search*				
Method	Venue	r=1	r=10	r=20	mAP	mINP	r=1	r=10	r=20	mAP	mINP
Zero-Pad [40]	ICCV17	14.80	54.12	71.33	15.95	-	20.58	68.38	85.79	26.92	-
HCML [44]	AAAI18	14.32	53.16	69.17	16.16	-	24.52	73.25	86.73	30.08	-
cmGAN [7]	IJCAI18	26.97	67.51	80.56	27.80	-	31.63	77.23	89.18	42.19	-
HSME [11]	AAAI19	20.68	32.74	77.95	23.12	-	-	-	-	-	-
AliGAN [33]	ICCV19	42.40	85.00	93.70	40.70	-	45.90	87.60	94.40	54.30	-
CMSP [39]	IJCV20	43.56	86.25	-	44.98	-	48.62	89.50	-	57.50	-
JSIA [34]	AAAI20	38.10	80.70	89.90	36.90	-	43.80	86.20	94.20	52.90	-
XIV [16]	AAAI20	49.92	89.79	95.96	50.73	-	-	-	-	-	-
MACE [43]	TIP20	51.64	87.25	94.44	50.11	-	57.35	93.02	97.47	64.79	-
MSR [9]	TIP20	37.35	83.40	93.34	38.11	-	39.64	89.29	97.66	50.88	-
DDAG [46]	ECCV20	54.75	90.39	95.81	53.02	-	61.02	94.06	98.41	67.98	-
Hi-CMD [6]	CVPR20	34.94	77.58	-	35.94	-	-	-	-	-	-
cm-SSFT [19] [1]	CVPR20	47.70	-	-	54.10	-	-	-	-	-	-
AGW [47]	TPAMI21	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23
MCLNet	**Ours**	**65.40**	**93.33**	**97.14**	**61.98**	**47.39**	**72.56**	**96.98**	**99.20**	**76.58**	**72.10**

[1] This paper reports a higher accuracy by using relation among all the gallery images. We use the results of single query for fair comparison.

Table 5. Comparison with the state-of-the-arts on RegDB dataset. Rank-k accuracy (%), mAP(%) and mINP (%) are reported.

Settings		*Visible to Infrared*					*Infrared to Visible*				
Method	Venue	r=1	r=10	r=20	mAP	mINP	r=1	r=10	r=20	mAP	mINP
Zero-Pad [40]	ICCV17	17.75	34.21	44.35	18.90	-	16.63	34.68	44.25	17.82	-
HCML [44]	AAAI18	24.44	47.53	56.78	20.08	-	21.70	45.02	55.58	22.24	-
HSME [11]	AAAI19	50.85	73.36	81.66	47.00	-	50.15	72.40	81.07	46.16	-
AliGAN [33]	ICCV19	57.90	-	-	53.60	-	56.30	-	-	53.40	-
CMSP [39]	IJCV20	65.07	83.71	-	64.50	-	-	-	-	-	-
JSIA [34]	AAAI20	48.50	-	-	48.90	-	-	-	-	-	-
XIV [16]	AAAI20	62.21	83.13	91.72	60.18	-	-	-	-	-	-
DG-VAE [26]	ACMMM20	72.97	86.89	-	71.78	-	-	-	-	-	-
HAT [48]	TIFS20	71.83	87.16	92.16	67.56	-	70.02	86.45	91.61	66.30	-
MSR [9]	TIP20	48.43	70.32	79.95	48.67	-	-	-	-	-	-
MACE [43]	TIP20	72.37	88.40	93.59	69.09	-	72.12	88.07	93.07	68.57	-
DDAG [46]	ECCV20	69.34	86.19	91.49	63.46	-	68.06	85.15	90.31	61.80	-
Hi-CMD [6]	CVPR20	70.93	86.39	-	66.04	-	-	-	-	-	-
AGW [47]	TPAMI21	70.05	86.21	91.55	66.37	50.19	70.49	87.12	91.84	65.90	51.24
MCLNet	**Ours**	**80.31**	**92.70**	**96.03**	**73.07**	**57.39**	**75.93**	**90.93**	**94.59**	**69.49**	**52.63**

expensive and space-expensive images generation, which reduces intermediate steps and avoids introducing additional noise. 2) Compared with the method using both global features and local features [46], our method significantly outperforms it by a large margin. Notably, our baseline model only computes global features. This would be beneficial for practical pedestrian retrieval deployment, while it has lower computational efficiency.

The experiments on RegDB dataset (Table 5) suggest that our proposed method is robust against different query settings. It performs well on both *visible-to-infrared* and *infrared-to-visible* settings by modality confusion learning. Since RegDB dataset is collected by a dual-camera system, we did not apply CCA on it. This learns better modality-invariant and modality-specific information.

5. Conclusion

In this paper, we propose a new cm-ReID baseline with a Modality Confusion Learning Network (MCLNet), which could learn modality-invariant features by minimizing inter-modality discrepancy while maximizing cross-modality similarity among instances. Different from other methods, MCLNet aims to confuse the two modalities by confusion learning mechanism. Meanwhile, we proposed an identity-aware and a camera-aware marginal center aggregation strategy for person ID and camera ID prediction, which can help the framework to extract the centralization features temperately. Extensive experiments validate the superior performance of the proposed method, as well as the effectiveness of each component of the framework.

Acknowledgement. This work is partially supported by the National Natural Science Foundation of China (62176188, 61902027) and CAAI-Huawei MindSpore Open Fund. The numerical calculations in this paper had been supported by the super-computing system in the Supercomputing Center of Wuhan University.

References

[1] Mindspore, https://www.mindspore.cn/, 2020. 6

[2] Sk Miraj Ahmed, Aske R. Lejbolle, Rameswar Panda, and

Amit K. Roy-Chowdhury. Camera on-boarding for person re-identification using hypothesis transfer learning. In *CVPR*, pages 12144–12153, 2020. 2

[3] Song Bai, Peng Tang, Philip HS Torr, and Longin Jan Latecki. Re-ranking via metric fusion for object retrieval and person re-identification. In *CVPR*, pages 740–749, 2019. 2

[4] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. Self-critical attention learning for person re-identification. In *ICCV*, pages 9637–9646, 2019. 2

[5] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *CVPR*, pages 8351–8361, 2019. 1, 2

[6] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *CVPR*, pages 10257–10266, 2020. 7, 8

[7] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, pages 677–683, 2018. 2, 3, 7, 8

[8] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, pages 994–1003, 2018. 2

[9] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE TIP*, 29:579–590, 2019. 1, 8

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 1

[11] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In *AAAI*, pages 8385–8392, 2019. 8

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

[13] Mengxi Jia, Yunpeng Zhai, Shijian Lu, Siwei Ma, and Jian Zhang. A similarity inference metric for rgb-infrared cross-modality person re-identification. *arXiv preprint arXiv:2007.01504*, 2020. 2

[14] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, pages 3143–3152, 2020. 2

[15] Kajal Kansal, AV Subramanyam, Zheng Wang, and Shin'ichi Satoh. Sdl: Spectrum-disentangled representation learning for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3422–3432, 2020. 1

[16] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *AAAI*, pages 4610–4617, 2020. 1, 2, 7, 8

[17] He Li, Mang Ye, and Bo Du. Weperson: Learning a generalized re-identification model from all-weather virtual data. In *ACMMM*, 2021. 1

[18] Fangyi Liu and Lei Zhang. View confusion feature learning for person re-identification. In *ICCV*, pages 6639–6648, 2019. 2, 4

[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, 8

[20] Chuanchen Luo, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Spectral feature transformation for person re-identification. In *CVPR*, pages 4976–4985, 2019. 2

[21] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, pages 0–0, 2019. 4

[22] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 7

[23] Hyeonjoon Moon and P Jonathon Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *Perception*, 30(3):303–321, 2001. 6

[24] 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

[25] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton Van Den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, pages 1846–1855, 2015. 2

[26] Nan Pu, Wei Chen, Yu Liu, Erwin M Bakker, and Michael S Lew. Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification. In *ACMMM*, pages 2149–2158, 2020. 2, 8

[27] Weijian Ruan, Mang Ye, Yi Wu, Wu Liu, Jun Chen, Chao Liang, Ge Li, and Chia-Wen Lin. Ticnet: A target-insight correlation network for object tracking. *IEEE TCYB*, 2021. 2

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 6

[29] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*, pages 608–617, 2019. 4

[30] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *CVPR*, pages 393–402, 2019. 2

[31] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 2

[32] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *CVPR*, pages 6449–6458, 2020. 2

[33] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality per-

son re-identification via joint pixel and feature alignment. In *ICCV*, pages 3623–3632, 2019. 2, 3, 7, 8

[34] Guan-An Wang, Tianzhu Zhang Yang, Jian Cheng, Jianlong Chang, Xu Liang, Zengguang Hou, et al. Cross-modality paired-images generation for rgb-infrared person re-identification. In *AAAI*, pages 12144–12151, 2020. 2, 3, 8

[35] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, pages 2275–2284, 2018. 2

[36] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. Person re-identification with cascaded pairwise convolutions. In *CVPR*, pages 1470–1478, 2018. 2

[37] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*, pages 618–626, 2019. 1, 2, 3, 6

[38] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515, 2016. 4

[39] Ancong Wu, Wei-Shi Zheng, Shaogang Gong, and Jianhuang Lai. Rgb-ir person re-identification by cross-modality similarity preservation. *IJCV*, pages 1–21, 2020. 8

[40] 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. 1, 2, 6, 8

[41] Dongming Wu, Mang Ye, Gaojie Lin, Xin Gao, and Jianbing Shen. Person re-identification by context-aware part attention and multi-head collaborative learning. *IEEE TIFS*, 2021. 2

[42] Bryan Ning Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. Second-order non-local attention networks for person re-identification. In *ICCV*, pages 3760–3769, 2019. 2

[43] Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE TIP*, 29:9387–9399, 2020. 2, 8

[44] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*, pages 7501–7508, 2018. 2, 6, 8

[45] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE TIFS*, 15:407–419, 2019. 1, 2

[46] 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*, 2020. 2, 7, 8

[47] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193*, 2020. 2, 3, 5, 6, 7, 8

[48] Mang Ye, Jianbing Shen, and Ling Shao. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE TIFS*, 2020. 8

[49] Shijie Yu, Shihua Li, Dapeng Chen, Rui Zhao, Junjie Yan, and Yu Qiao. Cocas: A large-scale clothes changing person dataset for re-identification. In *CVPR*, pages 3400–3409, 2020. 4

[50] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *CVPR*, pages 9021–9030, 2020. 2

[51] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *CVPR*, pages 10407–10416, 2020. 2

[52] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, pages 3186–3195, 2020. 1

[53] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *CVPR*, pages 8514–8522, 2019. 2

[54] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J Radke. Re-identification with consistent attentive siamese networks. In *CVPR*, pages 5735–5744, 2019. 2

[55] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM TOMM*, 14(1):1–20, 2017. 5

[56] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020. 6

[57] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, pages 5157–5166, 2018. 2

[58] Jiahuan Zhou, Bing Su, and Ying Wu. Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification. In *CVPR*, pages 2909–2918, 2020. 2

[59] Sanping Zhou, Fei Wang, Zeyi Huang, and Jinjun Wang. Discriminative feature learning with consistent attention regularization for person re-identification. In *ICCV*, pages 8040–8049, 2019. 2