Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification

Qiong Wu[1], Pingyang Dai[1]*, Jie Chen[2,7], Chia-Wen Lin[3], Yongjian Wu[4], Feiyue Huang[4],
Bineng Zhong[5], Rongrong Ji[1,6,7]
[1]Media Analytics and Computing Lab, Department of Artificial Intelligence,
School of Informatics, Xiamen University, 361005, China.
[2]School of Electronic and Computer Engineering, Peking University, China.
[3]National Tsing Hua University. [4]Tencent Youtu Lab. [5]Guangxi Normal University, China.
[6]Institute of Artificial Intelligence, Xiamen University. [7]Peng Cheng Laboratory, Shenzhen, China.
qiong@stu.xmu.edu.cn, pydai@xmu.edu.cn, chenj@pcl.ac.cn, cwlin@ee.nthu.edu.tw,
{littlekenwu, garyhuang}@tencent.com, bnzhong@gxnu.edu.cn, rrji@xmu.edu.cn

Abstract

Visible-infrared person re-identification (Re-ID) aims to match the pedestrian images of the same identity from different modalities. Existing works mainly focus on alleviating the modality discrepancy by aligning the distributions of features from different modalities. However, nuanced but discriminative information, such as glasses, shoes, and the length of clothes, has not been fully explored, especially in the infrared modality. Without discovering nuances, it is challenging to match pedestrians across modalities using modality alignment solely, which inevitably reduces feature distinctiveness. In this paper, we propose a joint Modality and Pattern Alignment Network (MPANet) to discover cross-modality nuances in different patterns for visible-infrared person Re-ID, which introduces a modality alleviation module and a pattern alignment module to jointly extract discriminative features. Specifically, we first propose a modality alleviation module to dislodge the modality information from the extracted feature maps. Then, We devise a pattern alignment module, which generates multiple pattern maps for the diverse patterns of a person, to discover nuances. Finally, we introduce a mutual mean learning fashion to alleviate the modality discrepancy and propose a center cluster loss to guide both identity learning and nuances discovering. Extensive experiments on the public SYSU-MM01 and RegDB datasets demonstrate the superiority of MPANet over state-of-the-arts.

1. Introduction

Person re-identification (Re-ID) [3] aims at matching individual pedestrian images in a query set to ones in a gallery set captured by different cameras. It is challenging due to the variations of viewpoints, body poses, illuminations, and backgrounds. Most existing person Re-ID methods [10, 20, 21, 22, 27, 32, 37, 39, 41, 42] focus on matching pedestrian images captured by visible cameras which can be formulated as a single-modality matching problem. However, these methods are not workable for images captured by visible surveillance cameras under poor illumination conditions (e.g., at night), from which it is difficult to extract discriminative information.

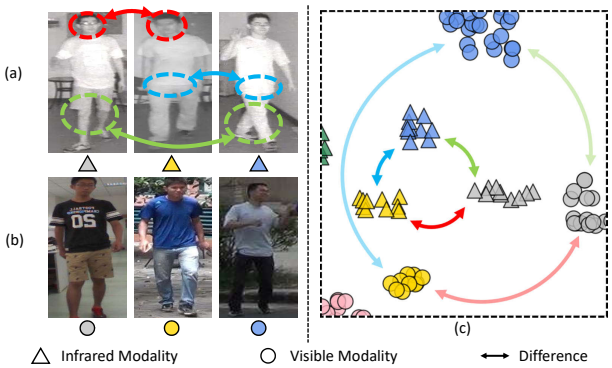

Figure 1. (a) Infrared and (b) visible pedestrian images, where the images in the same column are captured from the same identity. The difference among the visible identities is much more obvious than that among the infrared ones due to the limited information in the infrared modality. The nuances among different infrared images in different patterns provide a great number of differences which worth it discovering.

Cutting-edge surveillance systems are able to automatically switch from visible to infrared mode, which have accumulated a significant amount of cross-modality data. Re-ID problem in such a cross-modality setting thereby becomes extremely challenging, which is essentially a cross-modality retrieval problem. Compared to conventional per-

*Corresponding Author.

son Re-ID, new challenges arise from the modality by different spectrum cameras. As shown in Fig. 1, the infrared images of different identities in Fig. 1(a) are indiscernible, while the visible images in Fig. 1(b) are easy to distinguish. In addition, the appearances of a person inter modalities are completely different which is known as modality discrepancy. To perform visible-infrared person Re-ID, several methods [1, 4, 12, 29, 33] have been proposed, which aim to alleviate the modality discrepancy by aligning features or pixel distributions. Despite the encouraging achievement, the existing approaches still have limited ability in learning discriminative features across different modalities due to the efficient information buried in the infrared images that are not discovered. In cross-modality person Re-ID, the nuances in different image pairs arise in various patterns, such as the lengths of T-shirts and pants, the type of shoes, and wearing glasses or not. If this information is not well discovered, the discriminability of infrared features will be worse than the visible ones as shown in Fig. 1(c). Discovering nuances while alleviating modality discrepancy plays an important role in visible-infrared person Re-ID. Quite a few fine-grained person Re-ID approaches [15, 24, 27, 36, 40, 43] have been proposed recently, which mainly brings together identity classification, person auxiliary information into a framework to consider the details of a person. However, these methods require additional labeled priors, e.g., attributes, key points, and human parsing information, looking for certain parts and treat these parts equally rather than selecting them adaptively. Due to the lack of necessary information and the variations of modality, these methods fail to learn discriminative features in the cross-modality setting. Therefore, discovering nuances that are not fully exploited in existing methods can naturally improve the discrimination of features.

To fully explore nuanced information, we propose a novel cross-modality person Re-ID framework, termed joint Modality and Pattern Alignment Network (MPANet), which discovers cross-modality nuances while alleviating the modality discrepancy for visible-infrared person Re-ID. As shown in Fig. 2 , the proposed MPANet framework consists of two Modality Alleviation Modules (MAM) to alleviate modality discrepancy, a Pattern Alignment Module (PAM) to discover nuances in different patterns, and a mutual mean learning fashion to train the model with a center cluster loss and a cross-entropy loss for identity recognition. Specifically, MAM uses an instance normalization to alleviate the modality discrepancy while maintaining discriminative to the extend. By a light-weight generator, the pattern alignment module generates a group of pattern maps, which attend different patterns to discover nuances. The output of this module is obtained by concatenating both pattern features and the global feature. To discover nuances in an unsupervised manner, a region separation constraint is devised to ensure each pattern map attends to a different pattern. A center cluster loss is then proposed to reduce the distance among certain pattern features of the same identity while increasing the distance among the feature centers of different identities. We further apply two modality-specific classifiers to learn the identity of features from each modality and predict classification results of the same feature with them. Moreover, modality discrepancy is alleviated by reducing the distribution discrepancy between the predictions of the same image generated by different modality-specific classifiers in a mutual mean learning fashion. Finally, these two modules are cascaded and jointly optimized in an end-to-end manner. With the above work, the features extracted by MPANet are modality-invariant and can represent the nuances in different patterns.

Our main contributions are summarized below:

- We address the nuances discovery and modality discrepancy for visible-infrared person Re-ID in a unified framework. The former is not explored in the literature, while the latter is the key to matching the person across modalities.
- To discover the nuances and extract discriminative features, the pattern alignment module (PAM) is proposed to discover nuances in different patterns with a proposed center cluster loss and separation loss in an unsupervised manner.
- To alleviate the modality discrepancy while keeping the identity information, the modality alleviation module (MAM) is proposed which selectively applies instance normalization with the guide of a mutual mean learning manner.

2. Related Work

Visible-infrared Person Re-ID. Visible-infrared Person Re-ID has received increasing attention in recent years due to its effectiveness under poor illumination conditions. To address the challenge caused by modality discrepancy, many cross-modality person Re-ID approaches have been proposed. Wu *et al.* [33] proposed a deep zero-padding network learning features in a common space and construct the first large-scale visible-infrared dataset named SYSU-MM01. To constrain the intra-modality and inter-modality variations, an end-to-end dual-stream hyper-sphere manifold embedding model was proposed in [5]. In [35], a dual-path network with a bi-directional dual-constrained top-ranking loss was introduced to learn modality alignment feature representations. And *Ye et al.* also proposed a hierarchical cross-modality matching model that jointly optimizes the modality-specific and modality-shared metrics in [34]. DFE [4] was proposed to align the information both in region and modality. Some works are GAN-based

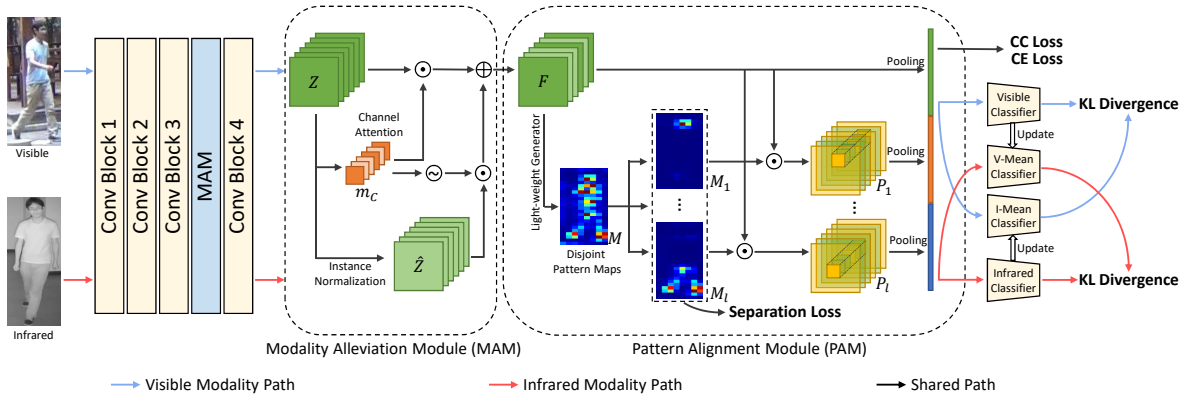

Figure 2. Framework of the proposed joint Modality and Pattern Alignment Network (MPANet). The Modality Alleviation Module (MAM) receives feature maps from the former block to extract modality-irrelevant feature maps. Subsequently, the Pattern Alignment Module (PAM) generates pattern maps to discover nuances in different patterns. A separation loss is proposed to ensure the pattern maps focus on the different patterns. Then the proposed center cluster loss instructs each pattern map to focus on a certain pattern and guide identity learning with the cross-entropy loss jointly. For guiding the network to alleviate the modality discrepancy, two modality-specific classifiers are applied with two corresponding mean classifiers in a mutual mean learning fashion.

approaches, cmGAN [1], D^2RL [29], AlignGAN [25] and JSIA-ReID [26]. cmGAN adopted generative adversarial training to map the features into a common space. D^2RL applied GANs to generate missing modality information extending the input of the feature extractor to four dimensions. Furthermore, AlignGAN and JSIA-ReID implemented pixel and feature dual-level alignment in a unified GAN framework. Similar, Li *et al.* [12] and cm-SSFT [14] generated a new modality between these two modalities to alleviate the modality discrepancy. Nevertheless, these approaches proposed to replenish the modality information or directly map the features into a common feature space. They mainly focus on alleviating the modality discrepancy while ignoring the effect of nuances, and it inevitably limits the boost of the performance.

Attention Mechanisms. The human visual system has an important property that humans selectively pay attention to salient parts of a series of glimpses to capture valuable information. Refer to the human visual system, there have been several attempts to adopt the attention mechanisms improving the performance of CNNs. Hu *et al.* introduced SENet [7] to exploit the dimension-wise relationship. They propose the Squeeze-and-Excitation module to apply attention mechanisms on the dimensions with global average-pooled features. Considering the relationship between any two positions, [28] proposed non-local neural network to capture the relationships among them. To broaden horizon, namely to make it can see 'what' and 'where' at the same time, CBAM [31] was proposed which exploits both spatial and dimension-wise attention. Following these methods, we propose the modality alleviation module (MAM) to protect identity by attention on channels while alleviating modality discrepancy. And we propose the pattern alignment module (PAM) to discover the nuances in different patterns.

Teacher-Student Models. In semi-supervised learning methods and knowledge distillation methods, teacher-student models play an important role. The critical idea of teacher-student models is to create consistent training supervision for each sample by collecting predictions from different models. Temporal ensembling [11] saved an average prediction in an exponential moving way for each sample as the supervisions of the unlabeled samples. To reduce the cost of saving predictions, Mean Teacher [23] temporally averaged model weights at different training iterations to create the supervisions for unlabeled samples. Different from the one-way transfer between a teacher and a student, deep mutual learning [38] was an ensemble of students who learn collaboratively and teach each other throughout the training process. Combines mutual learning and mean teaching, MMT [2] aimed to reduce the impact of noise from the pseudo label by using two mean teachers to generate soft labels for another two networks. Inspired by these methods, we make two modality-specific classifiers to predict features from both two modalities. In this way, the network is guided to extract modality-irrelevant features in this process by making the two classifiers with modality-specific knowledge but predict the same result.

3. Methodology

3.1. Problem Formulation

Let $\mathcal{V} = \{\mathbf{x}_v^i\}_{i=1}^{N_v}$ and $\mathcal{R} = \{\mathbf{x}_r^i\}_{i=1}^{N_r}$ respectively denote the visible images and infrared images in a cross-modality person Re-ID dataset, where N_v and N_r are the numbers of samples in each of the two modalities. There are totally $N = N_v + N_r$ samples in the dataset with the corresponding ground-truth label set $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^{N_p}$, where N_p is the number of identities. Given a query of a certain pedestrian, the

cross-modality person Re-ID aims to match the same person by finding a ranked list of images from another modality image set according to the similarity.

As shown in Fig. 2, the joint Modality and Pattern Alignment Network (MPANet) learns cross-modality representations to perform visible-infrared person Re-ID. MPANet adopts a pretrained one-stream CNN to extract feature maps from the visible and infrared modalities. The feature maps extracted by convolutional block 3 and 4 are respectively fed into the Modality Alleviation Module (MAM) that refines the feature maps to alleviate the modality discrepancy while preserving the identity discriminating ability of feature maps. To learn nuanced and discriminative features, the Pattern Alignment Module (PAM) generates pattern maps aiming to identify the nuances in different patterns of a person. These two modules are cascaded and jointly optimized by a mutual mean learning fashion to learn modality-irrelevant features and, meanwhile, are supervised by the cross-entropy and center cluster losses to learn identity-aware features for visible-infrared person Re-ID.

3.2. Modality Alleviation Module (MAM)

For an input image $\mathbf{x}$, we denote its feature map $\mathbf{Z} \in \mathbb{R}^{h\times w\times c}$ extracted by the convolutional block as the input of MAM, where h, w, c denote the height, width, and dimension of the feature map. To alleviate the modality discrepancy, we apply Instance Normalization (IN) which can reduce the discrepancy among instances [17]. Nevertheless, directly applying IN may damage identify information, thereby adversely affecting the Re-ID task.

In order to overcome these shortcomings, we apply channel attention-guided IN to alleviate modality discrepancy while preserving identity information:

$$\mathbf{F} = \mathbf{m}_C \odot \mathbf{Z} + (1 - \mathbf{m}_C) \odot \hat{\mathbf{Z}}, \tag{1}$$

where, $\mathbf{m}_C$ is the channel-wise mask indicating the identity-relevant channels, and $\hat{\mathbf{Z}}$ is the instance-normalized result of input $\mathbf{Z}$. Note that, the shape of $\mathbf{F}$ is the same as $\mathbf{Z}$.

Following SE-Net [8], we generate the dimension-wise mask $\mathbf{m}_C$ by

$$\mathbf{m}_C = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1 g(\mathbf{Z}))), \tag{2}$$

where $g(\cdot)$ denotes global average pooling, $\mathbf{W}_1 \in \mathbb{R}^{\frac{c}{r}\times c}$ and $\mathbf{W}_2 \in \mathbb{R}^{c\times\frac{c}{r}}$ are learnable parameters in the two bias-free fully-connected (FC) layers which are followed by ReLU activation function $\delta(\cdot)$ and sigmoid activation function $\sigma(\cdot)$. To balance the performance and complexity, the dimension reduction ratio $r = 16$ is used.

The parameter-free IN is defined as

$$\hat{\mathbf{Z}}_k = \mathrm{IN}(\mathbf{Z}_k) = \frac{\mathbf{Z}_k - \mathrm{E}[\mathbf{Z}_k]}{\sqrt{\mathrm{Var}[\mathbf{Z}_k] + \epsilon}}, \tag{3}$$

where $\mathbf{Z}_k \in \mathbb{R}^{h\times w}$ is the k-th dimension of feature map $\mathbf{Z}$, ϵ is used to avoid dividing-by-zero, the mean $\mathrm{E}[\cdot]$ and standard-deviation $\mathrm{Var}[\cdot]$ are calculated per-dimension.

3.3. Pattern Alignment Module (PAM)

To obtain discriminative features, PAM aims to discover the nuances in different patterns across identities. We split the feature map into l patterns with pattern maps $\mathbf{M} = [\mathbf{M}_1, \mathbf{M}_2, ..., \mathbf{M}_l] \in \mathbb{R}^{h\times w\times l}$. The maps are generated by a light-weight generator $\mathrm{A}(\cdot)$ as follows:

$$\mathbf{M} = \sigma(\mathrm{A}(\mathbf{F})), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid activation function, A is a convolution with kernel size 1. Note that, each of the pattern maps should pay attention to different patterns to discover the nuances included in them.

With these pattern maps, we can split the feature map $\mathbf{F}$ into l patterns as follows:

$$\mathbf{P}_k = \mathbf{M}_k \odot \mathbf{F}(k = 1, 2, ..., l), \tag{5}$$

where $\odot$ denotes element-wise multiplication.

Once the feature map is split into l patterns according to the pattern maps, the feature of the k-th pattern $\mathbf{p}_k = g(\mathbf{P}_k) \in \mathbb{R}^c$ is extracted by global average pooling $g(\cdot)$. Finally, the output feature $\mathbf{f} \in \mathbb{R}^{(l+1)c}$ of PAM can be represented by

$$\mathbf{f} = [{\mathbf{p}_1}^T, {\mathbf{p}_2}^T, ..., {\mathbf{p}_l}^T, g(\mathbf{F})^T]^T. \tag{6}$$

Generating pattern maps by the attention mechanism plays a key role in identifying cross-modality nuances. For person Re-ID, the pattern maps should cover diverse patterns of a person so that we can identify nuances involved in the diverse patterns. To ensure the pattern maps can capture different patterns, we apply the separation loss to force each map attending to different patterns. After resizing the mask $\mathbf{M}^{h\times w\times l}$ to $\mathbf{M}^{hw\times l}$, the separation loss is defined as

$$L_{sep} = \frac{2}{l(l-1)} \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} (\mathbf{M}^T\mathbf{M})_{ij}, \tag{7}$$

where $(\mathbf{M}^T\mathbf{M})_{ij}$ is the element of $\mathbf{M}^T\mathbf{M}$ on row i and column j. By minimizing the overlapping area between every two masks, the separation loss can supervise the pattern maps to learn features from diverse patterns.

3.4. Modality Learning (ML)

Given features $\mathbf{f}_v$ from the visible modality and $\mathbf{f}_r$ from the infrared modality, the modality-specific classifiers provide their predictions. These classifiers are trained with the

following cross-entropy loss in a supervised manner:

$$\begin{aligned}\mathcal{L}_{sid} = &-\frac{1}{n}\sum_{i=1}^{n}\log P(\mathbf{y}_v^i|\mathrm{C}_v(\mathbf{f}_v^i|\theta_v)) \\ &-\frac{1}{m}\sum_{j=1}^{m}\log P(\mathbf{y}_r^j|\mathrm{C}_r(\mathbf{f}_r^j|\theta_r)),\end{aligned} \tag{8}$$

where n and m respectively denote the numbers of visible and infrared images in the current batch, $\mathbf{y}_v^i$ and $\mathbf{y}_r^j$ respectively denote the corresponding label of $\mathbf{f}_v^i$ and $\mathbf{f}_r^j$, and $\mathrm{C}_v(\mathbf{f}_v^i|\theta_v)$ and $\mathrm{C}_r(\mathbf{f}_r^j|\theta_r)$ are predictions of the two classifiers with parameter θ_v and θ_r, respectively.

As the training images fed to each classifier come from a certain modality, the classifier learns the knowledge only from its corresponding modality. Thus, given a feature $\mathbf{f}$, no matter which modality it comes from, if two modality-specific classifiers provide the same prediction, it means this feature can be regarded as from both two modalities. In other words, the modality discrepancy is eliminated.

To this end, we impose a modality constraint based on Kullback-Leibler divergence as

$$\begin{aligned}\mathcal{L}_M = &\frac{1}{n}\sum_{i=1}^{n}\mathrm{C}_r(\mathbf{f}_v^i|\theta_r)\log\frac{\mathrm{C}_r(\mathbf{f}_v^i|\theta_r)}{\mathrm{C}_v(\mathbf{f}_v^i|\theta_v)} \\ &+\frac{1}{m}\sum_{j=1}^{m}\mathrm{C}_v(\mathbf{f}_r^j|\theta_v)\log\frac{\mathrm{C}_v(\mathbf{f}_r^j|\theta_v)}{\mathrm{C}_r(\mathbf{f}_r^j|\theta_r)}.\end{aligned} \tag{9}$$

This loss encourages the modality-specific classifiers to provide consistent predictions for the same-identity feature, no matter what modalities it comes from. However, training the model with Eq. (9) directly will make the predictions of the two classifiers become similar quickly since the classifiers learn the knowledge from another modality with Eq. (9), rather than learning modality-irrelevant features.

To address the above problem, we propose two mean classifiers with the same network structure as the modality-specific classifiers to provide predictions for samples from another modality. In this way, Eq. (9) can be modified as

$$\begin{aligned}\mathcal{L}_{MM} = &\frac{1}{n}\sum_{i=1}^{n}\mathrm{C}_r(\mathbf{f}_v^i|\mathrm{E}[\theta_r])\log\frac{\mathrm{C}_r(\mathbf{f}_v^i|\mathrm{E}[\theta_r])}{\mathrm{C}_v(\mathbf{f}_v^i|\theta_v)} \\ &+\frac{1}{m}\sum_{j=1}^{m}\mathrm{C}_v(\mathbf{f}_r^j|\mathrm{E}[\theta_v])\log\frac{\mathrm{C}_v(\mathbf{f}_r^j|\mathrm{E}[\theta_v])}{\mathrm{C}_r(\mathbf{f}_r^j|\theta_r)},\end{aligned} \tag{10}$$

where $\mathrm{E}[\theta_v]$ and $\mathrm{E}[\theta_r]$ denote the parameters of the two mean classifiers, respectively. These parameters are updated in a temporal average manner. Thus, at the t-th iteration, parameters $\mathrm{E}^{(t)}[\theta_v]$ and $\mathrm{E}^{(t)}[\theta_r]$ are calculated by

$$\begin{aligned}\mathrm{E}^{(t)}[\theta_v] &= (1-\alpha)\mathrm{E}^{(t-1)}[\theta_v] + \alpha\theta_v, \\ \mathrm{E}^{(t)}[\theta_r] &= (1-\alpha)\mathrm{E}^{(t-1)}[\theta_r] + \alpha\theta_r,\end{aligned} \tag{11}$$

where $\mathrm{E}^{(t)}[\theta]$ and $\mathrm{E}^{(t-1)}[\theta]$ respectively denote the parameters of mean classifiers in the current iteration and last iteration. The mean classifiers are initialized as $\mathrm{E}^{(0)}[\theta_v] = \theta_v$ and $\mathrm{E}^{(0)}[\theta_r] = \theta_r$. The parameter α is the updating ratio within the range of $(0, 1]$.

3.5. Objective Functions

Once the features have been extracted by the model, we train the model with the cross-entropy loss and center cluster loss. The following cross-entropy loss is imposed on classifier $\mathrm{C}(\cdot)$ to predict the identities:

$$\begin{aligned}\mathcal{L}_{id} = &-\frac{1}{n}\sum_{i=1}^{n}\log P(\mathbf{y}_v^i|\mathrm{C}(\mathbf{f}_v^i|\theta)) \\ &-\frac{1}{m}\sum_{j=1}^{m}\log P(\mathbf{y}_r^j|\mathrm{C}(\mathbf{f}_r^j|\theta)),\end{aligned} \tag{12}$$

where $\mathrm{C}(\mathbf{f}_v^i|\theta)$ and $\mathrm{C}(\mathbf{f}_r^j|\theta)$ are the identity predictions of $\mathbf{f}_v^i$ and $\mathbf{f}_r^j$ with a same classifier.

Furthermore, we propose the center cluster loss to learn the relationships among the identities and ensure each pattern map can always focus on a certain pattern as follows:

$$\begin{aligned}\mathcal{L}_{cc} = &\frac{1}{n+m}\sum_{i=1}^{n+m}||\mathbf{f}_i - \mathbf{h}_{\mathbf{y}_i}||_2 \\ &+\frac{2}{P(P-1)}\sum_{k=1}^{P-1}\sum_{j=k+1}^{P}[\rho - ||\mathbf{h}_{\mathbf{y}_k} - \mathbf{h}_{\mathbf{y}_j}||_2]_+,\end{aligned} \tag{13}$$

where $\mathbf{h}_{y_i}$ is the mean of features with label y_i in the current batch, P is the number of identities in the current batch and ρ is the least margin among the centers.

The center cluster loss aims at gathering the features to their center. Besides, the pattern features extracted from a certain intra-identity pattern will get close to each other. In this process, the model learns nuance information in an unsupervised manner. Meanwhile, the loss builds the relationship among classes directly rather than among samples, which bases the identity learning on the class-level and avoids increasing the modality discrepancy while pushing away different-identity samples.

3.6. Optimization

The total loss $\mathcal{L}$ of MPANet is defined as

$$\mathcal{L} = \mathcal{L}_{id} + \mathcal{L}_{cc} + \lambda_1\mathcal{L}_{sep} + \lambda_2\mathcal{L}_{sid} + \lambda_3\mathcal{L}_{MM}, \tag{14}$$

where λ_1, λ_2 and λ_3 are hype-parameters to balance the contributions of individual loss terms.

4. Experiments

4.1. Datasets and Experimental Setting

Datasets. We evaluate our method on two public datasets **SYSU-MM01** [33] and **RegDB** [16].

Table 1. Comparison of CMC (%) and *m*AP (%) performances with the state-of-the-art methods on **SYSU-MM01**

Method	All-Search								Indoor-Search							
	Single-Shot				Multi-Shot				Single-Shot				Multi-Shot			
	R1	R10	R20	*m*AP	R1	R10	R20	*m*AP	R1	R10	R20	*m*AP	R1	R10	R20	*m*AP
Two-stream [33]	11.65	47.99	65.50	12.85	16.33	58.35	74.46	8.03	15.60	61.18	81.02	21.49	22.49	72.22	88.61	13.92
One-stream [33]	12.04	49.68	66.74	13.67	16.26	58.14	75.05	8.59	16.94	63.55	82.10	22.95	22.62	71.74	87.82	15.04
Zero-Padding [33]	14.80	54.12	71.33	15.95	19.13	61.40	78.41	10.89	20.58	68.38	85.79	26.92	24.43	75.86	91.32	18.86
cmGAN [1]	26.97	67.51	80.56	27.80	31.49	72.74	85.01	22.27	31.63	77.23	89.18	42.19	37.00	80.94	92.11	32.76
D^2RL [29]	28.90	70.60	82.40	29.20	-	-	-	-	-	-	-	-	-	-	-	-
JSIA-ReID [26]	38.10	80.70	89.90	36.90	45.10	85.70	93.80	29.50	43.80	86.20	94.20	52.90	52.70	91.10	96.40	42.70
AlignGAN [25]	42.40	85.00	93.70	40.70	51.50	89.40	95.70	33.90	45.90	87.60	94.40	54.30	57.10	92.70	97.40	45.30
cm-SSFT(sq) [14]	47.70	-	-	54.10	-	-	-	-	57.40	-	-	59.10	-	-	-	-
DFE [4]	48.71	88.86	95.27	48.59	54.63	91.62	96.83	42.14	52.25	89.86	95.85	59.68	59.62	94.45	98.07	50.60
XIV-ReID [12]	49.92	89.79	95.96	50.73	-	-	-	-	-	-	-	-	-	-	-	-
CMM+CML [13]	51.80	92.72	97.71	51.21	56.27	94.08	98.12	43.39	54.98	94.38	99.41	63.70	60.42	96.88	99.50	53.52
SIM [9]	56.93	-	-	60.88	-	-	-	-	-	-	-	-	-	-	-	-
CoAL [30]	57.22	92.29	97.57	57.20	-	-	-	-	63.86	95.41	98.79	70.84	-	-	-	-
DG-VAE [18]	59.49	93.77	-	58.46	-	-	-	-	-	-	-	-	-	-	-	-
cm-SSFT [14]	61.60	89.20	93.90	63.20	63.40	91.20	95.70	62.00	70.50	94.90	97.70	72.60	73.00	96.30	99.10	72.40
MPANet (Ours)	**70.58**	**96.21**	**98.80**	**68.24**	**75.58**	**97.91**	**99.43**	**62.91**	**76.74**	**98.21**	**99.57**	**80.95**	**84.22**	**99.66**	**99.96**	**75.11**

- **SYSU-MM01** is a large-scale dataset collected by four visible cameras and two near-infrared ones, including both indoor and outdoor environments. The training set contains $22,258$ visible images and $11,909$ infrared ones involving 395 identities, while the query set and the gallery set contain $3,803$ infrared images and 301 ($3,010$) randomly sampled visible images from 96 identities for *single-shot* (*multi-shot*).
- **RegDB** is constructed by a pair of aligned cameras (one visible and one thermal). It contains $8,240$ images of 412 identities, each having 10 images from the visible camera and 10 images from the thermal one. The dataset is randomly split into two halves: the images of 206 identities for training and the rest also involving 206 identities for testing.

Evaluation metrics. To perform a fair comparison with existing methods, all experiments follow the common evaluation settings in existing cross-modality Re-ID methods. **SYSU-MM01** has two different evaluation settings: the *all-search* mode and *indoor-search* mode. In the *all-search* mode, the gallery set contains images from all the visible cameras, while in the *indoor-search* mode, the gallery set only contains images from the indoor visible cameras. Following [35], **RegDB** contains two test modes: use infrared images as query set and visible images as gallery set, and vice versa. For both datasets, the Cumulative Matching Characteristic (CMC) and mean Average Precision (*m*AP) metrics are adopted to evaluate the performance.

Implementation details. We implement our MPANet with PyTorch and train it on a single RTX2080Ti GPU. The mini-batch size is set to 128. For each mini-batch, we randomly sample 16 identities and 8 images for each identity. The model is optimized by using Adam with an initial learning rate of 3.5×10^{-4}, which decays at the $80th$ and $120th$ epoch with a decay factor of 0.1, and the weight decay is set to 5×10^{-4}. The total number of training epochs is set to 140. The center cluster loss margin ρ is set to 0.7. The update ratio α is set to 0.2. The hype-parameters λ_1, λ_2 and λ_3 are set to 0.5, 0.5 and 2.5, respectively.

The ResNet-50 [6] pre-trained on ImageNet is employed as the backbone, where the stride size of the last convolutional layer is set to 1. The classifier C_v, C_r, and C are implemented by a single FC layer without bias. Consider the aspect ratio of raw images, the input images are re-scaled to a fixed size of 384×128. In the training stage, the input images are randomly flipped and erased with 50% probability.

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

We compare our MPANet with state-of-the-art (SOTA) visible-infrared cross-modality person Re-ID approaches. The compared SOTAs include three base methods (Two-stream, One-stream and Zero-Padding) [33], three GAN-based methods (cmGAN [1], AlignGAN [25] and JSIA-ReID [26]), three methods by aligning the modality on a middle modality (XIV-ReID [12], cm-SSFT [14] and CMM+CML [13]), one similarity-based method (SIM [9]), one distribution alignment with image generation method (DG-VAE [18]), and three dual-level alignment methods (DFE [4], D^2RL [29] and CoAL [30]).

Comparisons on SYSU-MM01. The comparison results on **SYSU-MM01** are shown in Table 1. The proposed MPANet outperforms existing SOTAs by large margins. Specifically, MPANet achieves the Rank-1 accuracy of 70.58% and *m*AP of 68.24% in the *all-search* and *single-shot* mode, significantly improving the Rank-1 accuracy by 8.98% and *m*AP by 5.04% over the best SOTA cm-SSFT.

Table 2. Comparison of the CMC (%) and *m*AP (%) performances with SOTAs on **RegDB**

Method	infrared2visible		visible2infrared	
	Rank-1	*m*AP	Rank-1	*m*AP
Zero-Padding [33]	16.7	17.9	17.8	18.9
D^2RL [29]	-	-	43.4	44.1
JSIA-ReID [26]	48.1	48.9	48.5	49.3
AlignGAN [25]	56.3	53.4	57.9	53.6
CMM+CML [13]	59.8	60.9	-	-
XIV-ReID [12]	62.3	60.2	-	-
DFE [4]	68.0	66.7	70.2	69.2
cm-SSFT [14]	71.0	71.7	72.3	72.9
DG-VAE [18]	-	-	73.0	71.8
CoAL [30]	74.1	69.9	-	-
SIM [9]	75.2	78.3	74.7	75.2
MPANet(Ours)	**82.8**	**80.7**	**83.7**	**80.9**

Table 3. Ablation study in terms of CMC (%) and *m*AP (%) on **SYSU-MM01**

Method	SYSU-MM01			
	single-shot all-search			
	Rank-1	Rank-10	Rank-20	*m*AP
Baseline	54.50	88.55	94.69	51.84
B + CC	57.56	91.74	96.89	56.59
B + ML	58.55	90.01	94.75	55.06
B + ML + CC	64.24	93.70	97.48	61.39
B + ML + CC + PAM	68.27	93.97	97.84	64.98
B + ML + CC + PAM + MAM	70.58	96.21	98.80	68.24

When compared to the SOTAs in *indoor-search* and *multi-shot* mode, the performance margin between our MPANet and cm-SSFT are also significantly, e.g., the Rank-1 boost is 11.22%, and the *m*AP boost is 2.71%.

Comparisons on RegDB. We also evaluate MPANet on a small-scale dataset, **RegDB**, as shown in Table 2. Similar to the results on **SYSU-MM01**, MPANet consistently outperforms current SOTAs. Specifically, we achieve Rank-1 accuracy of 82.8% and *m*AP of 80.7% in *infrared2visible* mode, and Rank-1 accuracy of 83.7% and *m*AP of 80.9% in *visible2infrared* mode, significantly improving the Rank-1 by 7.6% and *m*AP by 2.4% in *infrared2visible* mode over the best SOTA SIM.

The above results demonstrate the outstanding performance of MPANet thanks to its ability in cross-modality nuances discovery for visible-infrared person Re-ID.

4.3. Ablation Study

In this section, we conduct an ablation experiment to evaluate the contribution of each module. The baseline method uses ResNet-50 as the backbone network followed by the BN neck and an FC layer as the classifier and trained with $\mathcal{L}_{id}$ in the same setting. The ablation experiment is conducted on **SYSU-MM01** in the *all-search single-shot* mode. To illustrate the contribution of each module or objective function, we add them into the model one by one.

As shown in Table 3, the effectiveness of each component is revealed. Compare with baseline, the center cluster (CC) loss and modality learning (ML) manner respectively improve the *m*AP accuracy by 4.75% and 3.22%. When these two objective functions work together to learn identity and alleviate modality discrepancy, the Rank-1 accuracy and *m*AP accuracy are significantly improved by 9.74% and 9.55%. Based on the above results, the PAM and MAM further improve the Rank-1 accuracy by 4.03% and 2.31%, respectively, with the guide of center cluster loss and mutual mean learning manner. The results demonstrate that each module plays a role effectively in alleviating modality discrepancy or improving discriminability.

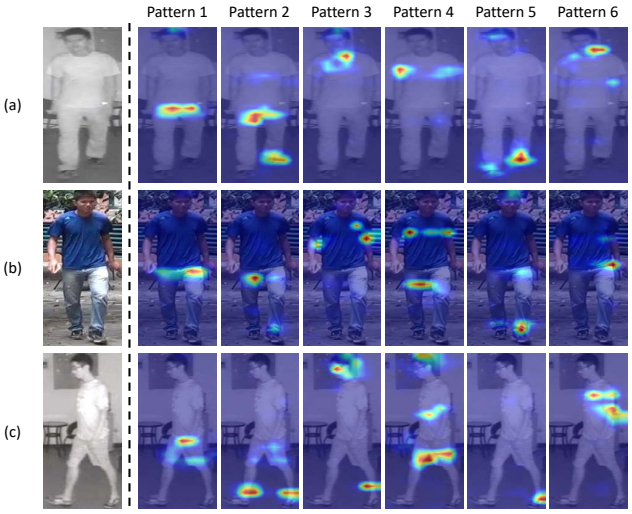

Figure 3. Visualization result of PAM on **SYSU-MM01**. For each row, we show an input image and six patterns corresponding to six pattern maps. The patterns in the same column are extracted by the same channel of pattern maps. (a) and (b) have the same identity, and (c) has a different identity.

4.4. Discussions

Attention to Patterns. One of the keys to visible-infrared person Re-ID is to improve the discriminability of features. To further illustrate the effectiveness of PAM which can attend diverse patterns for discovering the nuances, we visualize the pixel-wise pattern maps learned by PAM. In general, each pattern map should have a corresponding pattern of interest. We apply Grad-Cam [19] to visualize these areas by highlighting them on the image. Fig. 3 illustrates individual attention patterns for the three pedestrian images of two identities, where the k-th column is the visualization result of $\mathbf{M_k} \odot \mathbf{F}$. We can observe that every pattern map generated by PAM focuses on a certain pattern different from the others without the effect of modality and pose. The visualization demonstrates the importance of nuances in this task, and the PAM works well on it.

Visualized Distributions. To illustrate the impact of MPANet on alleviating modality discrepancy and on increasing discriminability, we randomly select 10 identities

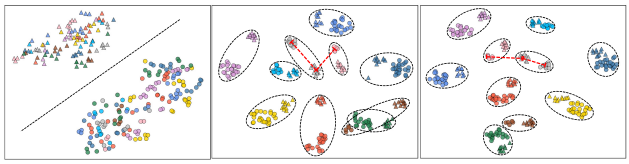

Figure 4. Visualization of learned features, where each color represents an identity in the testing set. The circles and triangles indicate the features extracted from the visible and infrared modalities. (a) Features extracted by the baseline pre-trained on ImageNet; (b) Features extracted by the baseline; (c) Features extracted by our MPANet. It is obvious that MPANet better alleviates the modality discrepancy and improves the discriminability.

from the testing set to visualize the distributions of learned features by t-SNE in Fig. 4, where each distinct color represents an identity while the circles and triangles indicating the features of visible and infrared images, respectively. As shown in Fig. 4(a), the initial features have significant modality discrepancy and it is difficult to match the same person across the modalities. In Fig. 4(b), although most features extracted by the baseline can be clustered well, the intra-identity modality discrepancy remains obvious. Moreover, the infrared images of some identities (such as light blue, pink, and gray) are gathered. Thus some infrared images (for example, the gray triangles) may match the wrong visible ones. This motivates the need for discovering the nuances while alleviating the modality discrepancy. In contrast, as shown in Fig. 4(c), the learned features of different modalities with MPANet are well grouped by identity. Furthermore, improving the discriminability of features by discovering nuances, especially for the features from infrared images achieves a significant improvement.

Effect of Attention Mechanisms. The attention mechanisms play an important role in the proposed MPANet. To demonstrate the advantages of the proposed MAM and PAM, we compare them with existing attention methods, including SE block [8], CBAM [31] and Non-local block [28], and replace MAM with Instance Normalization to directly alleviate the modality discrepancy. As shown in Table 4, '$A \rightarrow B$' means replace A with B while keeping the other modules. For a fair comparison, the different attention mechanism modules which work on channels or spatial are used to replace the modules that work in the same way.

When MAM is replaced by the SE block, it selects the channels including the identity relevant information but ignores alleviating the modality discrepancy. Although replacing MAM with Instance Normalization can alleviate the modality discrepancy but the discriminability will be harmed at the same time. Compare with MPANet, these two methods drop the Rank-1 accuracy by 4.75% and 9.26%. When PAM is replaced by Non-local Block, the model mines the relationship in the spatial domain rather than discovers the nuances, and the Rank-1 accuracy and *m*AP accuracy drop by 6.71% and 7.85%, respectively. Finally, we replace MAM and PAM with CBAM, the Rank-1 accuracy and *m*AP accuracy drop by 6.86% and 6.70%, respectively. The result indicates that MAM and PAM outperform the other attention methods for visible-infrared person Re-ID.

Table 4. Performance comparison with other attention mechanisms in terms of CMC (%) and *m*AP (%) on **SYSU-MM01**

Method	SYSU-MM01			
	single-shot all-search			
	Rank-1	Rank-10	Rank-20	*m*AP
MPANet	70.58	96.21	98.80	68.24
MAM $\rightarrow$ SE	65.83	93.92	97.62	63.29
MAM $\rightarrow$ IN	61.32	91.36	96.11	58.35
PAM $\rightarrow$ NL	63.87	93.20	97.36	60.39
MAM + PAM $\rightarrow$ CBAM	63.72	93.48	97.26	61.54

5. Conclusion

In this paper, we proposed the joint modality and pattern alignment network, termed MPANet, to discover cross-modality nuances for visible-infrared person Re-ID. Our method aims to both alleviate the modality discrepancy and discover the nuances in different patterns which are the keys to solving this task. To this end, the proposed MPANet focuses on extracting modality-irrelevant features that particularly attend to the identity-aware nuances among identities. Specifically, MPANet first employs two Modality Alleviation Modules (MAM) which selectively apply instance normalization to alleviate the modality discrepancy while preserving the identity information. Then, in the Pattern Alignment Module (PAM), the feature maps are split into multiple patterns according to the pattern maps to extract features from each pattern to discover the nuances. We optimize the MPANet in an end-to-end manner, where a mutual mean learning fashion that works as a cross-modality discrepancy constraint. And the center cluster loss is proposed to learn nuance information and guides identity learning on the class-level. Experiment results on two public datasets SYSU-MM01 and RegDB amply proves essential to discover cross-modality nuances in cross-modality retrieval problem and demonstrate the effectiveness of MPANet for visible-infrared person Re-ID.

Acknowledgements. This work is supported by the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No.U1705262, No.62072386, No.62072387, No.62072389, No.62002305, No.61772443, No.61802324, No.61702136, No.61972217 and No.62081360152), Guangdong Basic and Applied Basic Research FoundationNo.2019B1515120049) and Guangdong Science and Technology Department (No.2020B1111340056).

References

[1] 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, 6

[2] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 3

[3] Shaogang Gong, Marco Cristani, Chen Change Loy, and Timothy M. Hospedales. The re-identification challenge. In *Person Re-Identification*, pages 1–20. 2014. 1

[4] Yi Hao, Nannan Wang, Xinbo Gao, Jie Li, and Xiaoyu Wang. Dual-alignment feature embedding for cross-modality person re-identification. In *ACMMM*, pages 57–65, 2019. 2, 6, 7

[5] 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. 2

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

[7] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *PAMI*, pages 2011–2023, 2020. 3

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 4, 8

[9] Mengxi Jia, Yunpeng Zhai, Shijian Lu, Siwei Ma, and Jian Zhang. A similarity inference metric for rgb-infrared cross-modality person re-identification. In Christian Bessiere, editor, *IJCAI*, pages 1026–1032, 2020. 6, 7

[10] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, pages 3140–3149, 2020. 1

[11] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 3

[12] 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. 2, 3, 6, 7

[13] Yongguo Ling, Zhun Zhong, Zhiming Luo, Paolo Rota, Shaozi Li, and Nicu Sebe. Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *ACMMM*, pages 889–897, 2020. 6, 7

[14] 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 13376–13386, 2020. 3, 6, 7

[15] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 542–551. IEEE, 2019. 2

[16] 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. 5

[17] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*. 4

[18] 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 Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *ACMMM*, pages 2149–2158, 2020. 6, 7

[19] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 7

[20] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *CVPR*, pages 719–728, 2019. 1

[21] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6397–6406, 2020. 1

[22] 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 501–518, 2018. 1

[23] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *ICLR*, 2017. 3

[24] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, pages 7134–7143, 2019. 2

[25] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *ICCV*, pages 3622–3631, 2019. 3, 6, 7

[26] Guan'an Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *AAAI*, pages 12144–12151, 2020. 3, 6, 7

[27] 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. IEEE Computer Society, 2018. 1, 2

[28] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 3, 8

[29] 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. 2, 3, 6, 7

[30] Xing Wei, Diangang Li, Xiaopeng Hong, Wei Ke, and Yihong Gong. Co-attentive lifting for infrared-visible person re-identification. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *ACMMM*, pages 1028–1037, 2020. 6, 7

[31] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In *ECCV*, pages 3–19, 2018. 3, 8

[32] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Unsupervised person re-identification by camera-aware similarity consistency learning. In *ICCV*, pages 6921–6930. IEEE, 2019. 1

[33] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5390–5399, 2017. 2, 5, 6, 7

[34] 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

[35] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C. Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, pages 1092–1099, 2018. 2, 6

[36] Jiahang Yin, Ancong Wu, and Wei-Shi Zheng. Fine-grained person re-identification. *IJCV*, 128(6):1654–1672, 2020. 2

[37] Hong-Xing Yu and Wei-Shi Zheng. Weakly supervised discriminative feature learning with state information for person identification. In *CVPR*, pages 5527–5537. IEEE, 2020. 1

[38] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018. 3

[39] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, pages 3183–3192, 2020. 1

[40] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-Sheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *CVPR*, pages 4913–4922, 2019. 2

[41] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 2138–2147, 2019. 1

[42] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *ECCV*, pages 176–192, 2018. 1

[43] Qinqin Zhou, Bineng Zhong, Xiangyuan Lan, Gan Sun, Yulun Zhang, Baochang Zhang, and Rongrong Ji. Fine-grained spatial alignment model for person re-identification with focal triplet loss. *TIP*, 29:7578–7589, 2020. 2