Visible-Infrared Person Re-Identification via Patch-Mixed Cross-Modality Learning

Zhihao Qian[a], Yutian Lin[a,*], Bo Du[a]

[a]*the School of Computer Science, Luojia laboratory, Wuhan University, Wuhan, 430072, China*

Abstract

Visible-infrared person re-identification (VI-ReID) aims to retrieve images of the same pedestrian from different modalities, where the challenges lie in the significant modality discrepancy. To alleviate the modality gap, recent methods generate intermediate images by GANs, grayscaling, or mixup strategies. However, these methods could introduce extra data distribution, and the semantic correspondence between the two modalities is not well learned. In this paper, we propose a Patch-Mixed Cross-Modality framework (PMCM), where two images of the same person from two modalities are split into patches and stitched into a new one for model learning. A part-alignment loss is introduced to regularize representation learning, and a patch-mixed modality learning loss is proposed to align between the modalities. In this way, the model learns to recognize a person through patches of different styles, thereby the modality semantic correspondence can be inferred. In addition, with the flexible image generation strategy, the patch-mixed images

*Corresponding author

freely adjust the ratio of different modality patches, which could further alleviate the modality imbalance problem. On two VI-ReID datasets, we report new state-of-the-art performance with the proposed method.
Keywords: Visible-Infrared Person Re-Identification, Patch-Mix

1. Introduction

Visible-infrared person re-identification (VI-ReID) [1] aims to match a target person between the RGB visible cameras and low-light infrared (IR) cameras. The task is increasing research interests [2, 3, 4, 5] because of its great value in the practical 24-hour surveillance system. The main challenge of VI-ReID lies in modality discrepancy, where different wavelengths bring significantly different visual appearances (*e.g.,* color, texture).

Typically, there are three main types of methods: 1) the representation learning based methods [6, 7], where networks are designed to learn discriminative features in a modality-shared space; 2) the metric learning based methods [8, 9], where loss functions are designed to bridge the modality gap; 3) the modality-transform based methods [10, 11], which transform modalities into each other for style consistency. However, these methods try to handle the large modality discrepancy directly, while it is hard to align heterogeneous modalities without considering the correspondence between IR and RGB images.

To handle the above issue, recent works generate a third modality to assist cross-modality learning. Among them, a branch of works constructs

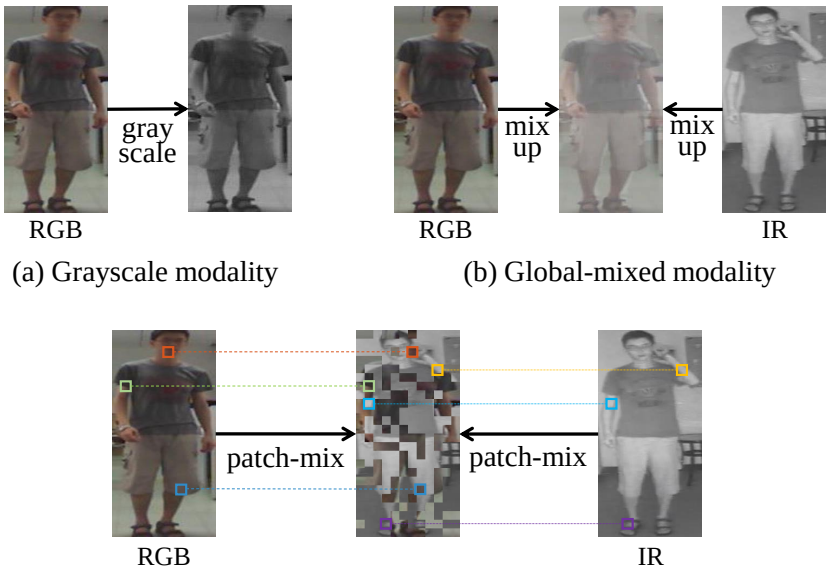


Figure 1: Different methods of generating the intermediate modality. (a) grayscale images are generated by visible images, (b) a mixed image is a global mixture of corresponding RGB and IR images, and (c) our proposed patch-mixed image, where each patch is either from RGB or IR, helps to infer the semantic corresponds between the two modalities and relieves the modality imbalance problem.

the new modality upon only one modality. For example, [12] generates the third auxiliary modality by transforming the visible images to one-channel images and then reconstructs three-channel images. In [13], grayscale images are generated by visible images and are utilized to enhance the robustness against color variations. However, as shown in Fig. 1 (a), these new images are generated through only one modality, where the information from the IR images is ignored. Another branch of works [14, 15] adopts mixup [16] strategy to generate intermediate modality images between RGB and infrared images, achieving promising results. However, as shown in Fig. 1 (b), it in-

troduces a different data distribution, with each pixel is neither RGB nor IR, which may improve the generalization ability but the semantic correspondence between the two modalities is chaotic that the person may appear as a double shadow with two heads.

In this paper, we propose a Patch-Mixed Cross-Modality framework (PMCM), which leverages modality discrepancy by learning with a new patch-mixed modality. As shown in Fig. 1 (c), the patch-mixed image is generated by blending a person's image of two modalities in the patch level where the data distribution within each patch is consistent with the original modality. The patch-mixed modality benefits cross-modality learning in two aspects: 1) The semantic correspondence between two modalities could be inferred by recognizing a patch-mixed image. For example, the network will learn that the hair in the IR patch and the face in the RGB patch together describe a person's appearance. Thereby, the model learns to deal with the two modalities in the same way, and the modality gap is reduced. 2) The patch-mixed modality also relieves the modality imbalance problem. As there usually are more images captured by daytime cameras, the data of IR and RGB images are imbalanced. With the flexible image generation approach, the proportion of different modalities can be adjusted freely to produce images with more IR information or RGB information. Therefore, the distribution of training samples is modified, which achieves an effect similar to over-sampling.

As a minor contribution, we adopted an improved center-to-center loss to directly reduce the modality gap by aligning identity centers between RGB,

IR and our patch-mixed modality. To regularize representation learning of part features, we take full advantage of the association between the part features and the global feature. Thus, a part alignment loss is proposed to constrain the consistency of part and global prediction distributions. With the part-based learning strategy, the discriminative part features are explored, which benefits the global feature learning in return. Besides, considering the shared information between the new modality and the other two modalities, we propose a patch-mixed modality learning loss to enhance the modality invariance learning by aligning the distribution of the prediction logits.

The main contributions of our work can be summarized as follows:

- We propose a novel patch-mixed cross-modality learning framework for the VI-ReID task, which effectively encourages the model to treat the RGB and IR images in the same way and alleviate the modality imbalance problem.

- We consider different constraints to further enhance the learned model. A part-alignment loss is proposed to constrain the consistency of part and global prediction distributions for more discriminative representation. A patch-mixed modality learning loss is proposed to align the new modality with the other two modalities.

- Experimental results show that our method outperforms other methods on two VI-ReID datasets by a large margin, and the data imbalance problem is effectively alleviated.

2. Related Work

2.1. Visible-Infrared Person Re-identification

VI-ReID aims to match persons of different modalities, which faces the challenge of large intra-modality variation and inter-modality discrepancy. Wu *et al.* [17] was the first to define the task, which proposed a deep zero-padding method along with a large-scale VI-ReID dataset named SYSU-MM01.

Following that, researchers propose to learn modality-specific and modality-shared feature representations by designing networks or loss functions. Ling *et al.* [8] propose a Multi-Constraint similarity learning method that jointly considers the cross-modality relationships from three different aspects. Sun *et al.* [18] performs pixel-to-pixel dense alignment acting on the intermediate representations. Huang *et al.* [6] makes use of both modality shared appearance features and modality-invariant relation features to boost performance Huang *et al.* [19] take the initiative to investigate the importance and strategy of exploiting person body information. Wan *et al.* [20] explicitly utilizes body topology to jointly achieve semantic- and structural-level alignment.

On the other hand, another branch of work aims to bridge the modality discrepancy by transforming the images from one modality to the other by generative adversarial networks (GANs). Choi *et al.* [21] propose an effective generator to extract pose-invariant and illumination-invariant features. Zhao *et al.* [22] learns the color-irrelevant features and aligns the identity-level feature distributions. Zhang *et al.* [23] compensates for the missing

modality-specific information from the other modality in the feature level.

2.2. VI-ReID with Intermediate Modality Images

To further reduce the modality gap, researchers propose to construct a third modality to assist shared feature space learning. In [12], a third auxiliary modality is generated by transforming the visible images to one-channel images and then reconstructing three-channel images. In [13], grayscale images are generated by visible images and are utilized to enhance the robustness against color variations. Following that, [11] transforms both of the two modalities into the grayscale for modality alignment to reduce the modality discrepancy. However, in these methods, the third modality is generated upon only one modality, without considering the relationship between IR and RGB images.

Recently, some methods have proposed to generate intermediate modality images between RGB and infrared images, which achieved promising results. In [24], inspired by mixup [16], a linear interpolation is performed of two images from different modalities for the same identities to generate mixed images. Similarly, in [14], mixed modality images are generated with a dynamic mixup ratio learned by a deep reinforcement learning framework. In [15], a syncretic modality collaborative learning model is designed, where shallow representation is mixed. Lu *et al.* [10] proposes an intermediate modality generation module involves CutMix [25] to better integrate features from visible and infrared modalities.

Compared with these extra-modality learning methods, our Patch-Mix strategy generates the third modality by mixing the RGB and IR modalities at the raw pixel level, which helps to learn a semantic alignment of two modalities by a unified input.

2.3. Modality Imbalanced Learning

Current research on imbalanced data focuses on the class imbalance problem and introduces two main strategies: re-sampling and re-weighting. Re-sampling like [26, 27] over-sample classes with few samples and under-sample classes with many samples. Re-weighting like [28, 29] adaptively adjusts the weights of different classes in the loss function. Liu *et al.* [30] first notice the unique data imbalance problem in cross-modality tasks and name it Modality Imbalance, which refers to the situation that one modality contains more samples than the other modality. To address the problem, they borrow the idea of re-weighting and allowing independent augmentation for a specified modality. Different from their work, our PMCM alleviates this problem by adjusting the ratio of patch-mix, where more patches of one modality can be contained for data balance.

3. Proposed Method

In this paper, we aim to learn modality-invariant representations by an intermediate patch-mixed modality, where cross-modality retrieval can be achieved. The overview of the proposed method is shown in Fig. 2, where

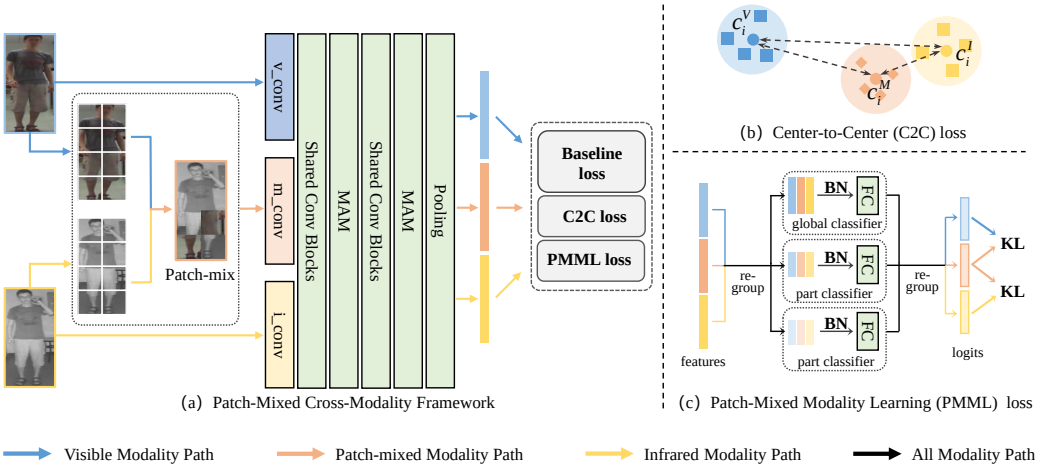


Figure 2: (a) Framework of the proposed PMCM. The patch-mixed image and the original images are together fed into the backbone network to extract features. After pooling, the obtained features are jointly optimized by the baseline loss, center-to-center loss, and patch-mixed modality learning loss. (b) Center-to-center (C2C) loss attempts to reduce the distance between the identity centers of any two modalities. (c) Patch-Mixed Modality Learning (PMML) loss aims to align the prediction distributions of the patch-mixed modality with that of the other two modalities, where global and local features are considered.

RGB, IR, and patch-mixed images are fed into the network, optimized by the baseline losses, center-to-center loss, and patch-mixed modality learning loss.

Following, we introduce our method in detail. We begin with the baseline method for this task. Then we illustrate the modality center alignment and patch-mixed cross-modality framework.

3.1. Baseline Method

3.1.1. Baseline

We adopt a two-stream network as our baseline and sample the same number of RGB and IR images in a mini-batch. We introduce the ResNet-50

as the backbone, where the first convolution blocks are modality-unique to learn modality-invariant low-level features, and the others are weight-shared to capture discriminative features. Besides, we alter channel-wise attention to part-wise attention in MAMs proposed by Wu *et al.* [31] and insert them into the backbone to extract modality-irrelevant feature maps. After the backbone, we only adopt global average pooling to the obtained feature maps to attain global features and then use a batch normalization layer to get the query representations. At last, the loss function for the baseline is formulated as follows:

$$\mathcal{L}_{base\_global} = \mathcal{L}_{id,g} + \mathcal{L}_{tri} + \lambda_1 \mathcal{L}_{s2s,g}, \tag{1}$$

where $\mathcal{L}_{id,g}$ is the cross-entropy loss of the global features after a classifier, $\mathcal{L}_{tri}$ is the hard triplet loss [32], λ_1 denotes the weight of $\mathcal{L}_{s2s,g}$, which is the sample-to-sample loss [11], attempting to pull close the sample features of different modalities with the same identity. Specifically, given the global features of two modalities, RGB and IR, the sample-to-sample loss is formulated as:

$$\mathcal{L}_{s2s} = \frac{1}{N} \sum_{i=1}^{N} mean[F(f_i^V) - F(f_i^I)], \tag{2}$$

where N is the number of paired samples in a mini-batch, f_i^V and f_i^I are the features of the i-th sample of RGB and IR, respectively. $F(\cdot)$ is a network with two fully-connected layers.

3.1.2. Baseline with Part-based Learning

Inspired by PCB [33], recent VI-ReID works [11, 22, 31] learn part-based features to enhance global discriminative representation learning and achieve promising performance. In this work, we also exploit horizontal stripes to obtain part features. Similarly, cross-entropy loss and sample-to-sample loss are adopted to address the part-based features:

$$\mathcal{L}_{base\_part} = \mathcal{L}_{id,p} + \mathcal{L}_{s2s,p}. \tag{3}$$

where $\mathcal{L}_{id,p}$ is the cross-entropy loss of all the part features and $\mathcal{L}_{s2s,p}$ is the sample-to-sample loss of all the part features.

In addition, since the global feature and part features both describe the same identity, we hope that the output distribution of part features could be similar to the global feature. Therefore, we calculate the KL divergence of the two output distributions as the regularization term, to learn more generalized part features. In this way, the prediction of global and each part features are supposed to be consistent. Given part and global features, the part alignment loss can be formulated as:

$$\mathcal{L}_{part\_align} = \sum_{i=1}^{N} \sum_{k=1}^{P} C_g(f_i^g) \log \frac{C_g(f_i^g)}{C_{p_k}(f_i^{p_k})}, \tag{4}$$

where P is the number of parts, $f_i^{p_k}$ denotes the feature of the k-th part feature of the i-th identity, $C_g(\cdot)$ and $C_{p_k}(\cdot)$ are the classifier of global features and that of the k-th part features.

Besides, in order to speed up the convergence of the model, we reduce the weight of losses involving part features in the early stage of training, which may cause relatively large interference. Therefore we set a weight μ to losses involving part features, which linearly increases with epochs and reaches its maximum at some point. In this way, the total loss of the part-based baseline is formulated as follows:

$$\mathcal{L}_{base} = \mathcal{L}_{base\_global} + \mu(\mathcal{L}_{base\_part} + \mathcal{L}_{part\_align}). \tag{5}$$

3.2. Modality Center Alignment

To further reduce the cross-modality variance, we consider directly aligning the identity centers between any two modalities. We adopt a global center-to-center loss similar to previous work [8], where the distance between each center of the same identity from different modalities is minimized as shown in Fig. 2(b). We obtain global centers of all training data by maintaining a memory bank instead of calculating the centers only in a mini-batch. Thus our global centers are more robust and tolerant to the mini-batch calculation bias caused by insufficient samples.

Given the global features of RGB images and IR images, the global center-to-center loss can be defined as:

$$\mathcal{L}_{c2c,g} = \frac{1}{Y}\sum_{i=1}^{Y}\left\|m_i^V - m_i^I\right\|^2, \tag{6}$$

where m_i^V and m_i^I respectively denote the memory banks of the center of the

i-th identity in the RGB and IR modality, which are updated every mini-batch, and Y is the number of identities.

Similarly, we consider the center relations between part features and use μ to balance the global and part losses. The final center-to-center loss is calculated as:

$$\mathcal{L}_{c2c} = \lambda_2 \mathcal{L}_{c2c,g} + \mu \lambda_3 \mathcal{L}_{c2c,p} \tag{7}$$

where λ_2 and λ_3 are the weights of $\mathcal{L}_{c2c,g}$ and $\mathcal{L}_{c2c,p}$.

In addition, there are also some differences in the optimization strategy. Since the stored features in the memory are inconsistent with those trained in the current training batch, we set a threshold epoch to delay the optimization of this loss until the model gets stable.

3.3. Patch-Mixed Cross-Modality Learning

To deal with the modality variance, recent VI-ReID works [13, 14, 24] construct intermediate modalities and achieve promising performance. Different from these grayscale or mixup-based methods, we propose a novel patch-mixed intermediate modality, where each patch is from the original IR and RGB images, having the same data distribution of the original data.

3.3.1. Patch-Mix strategy

Given an infrared image x^I and a visible image x^V, a patch-mixed image x^M is generated, where each patch $x^M(i,j)$ is formed by either $x^V(i,j)$ or $x^I(i,j)$. Here, i and j denote the patch index of image length and width, respectively. We set the probability of choosing an RGB patch to be p, and

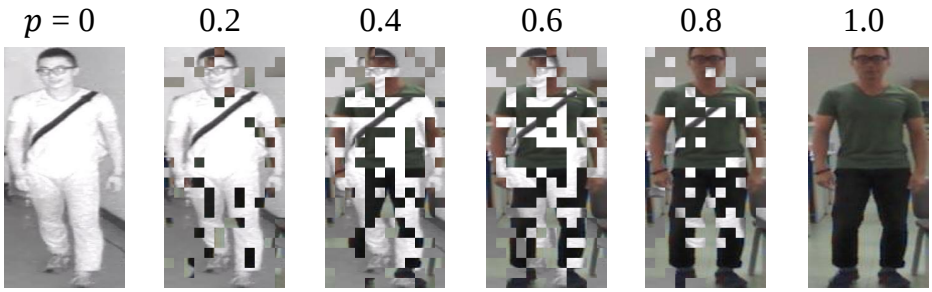


Figure 3: Patch-mixed images with different mix ratios p. When $p = 0$, the image is composed of only an infrared image. As p increases, more visible patches are adopted.

then the probability of choosing an infrared patch is $1 - p$. As shown in Fig. 3, the ratio p balances the information from two modalities.

The patch mixed images encourage cross-modality learning from two aspects. First, by learning from adjacent RGB and IR patches, which may represent the same semantic part, the model is encouraged to recognize the same semantic part through two modality patches. Therefore, the modality variance is bridged and the generalization ability is improved. Second, the patch-mixed images help to reduce the modality imbalance. Usually, images captured by daytime cameras are more than those of nighttime cameras, so the data of IR and RGB images are typically imbalanced. As shown in Fig. 3, with an adjustable ratio p, images with more IR information or RGB information can be generated freely according to the data distribution.

3.3.2. Patch-Mixed Modality Learning (PMML)

Considering that the mixed modality contains internal information in both of the two modalities, we align it with the other two modalities.

As shown in Fig. 2 (c), we input three training samples in three modali-

ties with the same identity into the network and obtain their output distribution. The KL divergence of the two output distributions is calculated to constrain the learning of patch-mixed modality. Since the logits are grouped by global and part relationship after the classifiers, we first regroup the logits by modality and then align between the mixed modality and the two original modalities.

Given part and global features, the part alignment loss can be formulated as:

$$\begin{aligned}\mathcal{L}_{pmml}^{M,V} = &\sum_{i=1}^{K} C_g(f_i^{V,g}) \log \frac{C_g(f_i^{V,g})}{C_{p_k}(f_i^{M,g})} \\ &+ \sum_{i=1}^{K}\sum_{k=1}^{P} C_{p_k}(f_i^{V,p_k}) \log \frac{C_{p_k}(f_i^{V,p_k})}{C_{p_k}(f_i^{M,p_k})}\end{aligned} \tag{8}$$

where $f_i^{p_k}$ denotes the feature of the k-th part feature of the i-th identity, $C_g(\cdot)$ and $C_{p_k}(\cdot)$ are the classifier of global features and that of the k-th part features.

Similarly, the alignment between the patch-mixed images and infrared images is calculated. The total patch-mixed modality learning loss is defined as:

$$\mathcal{L}_{pmml} = p\mathcal{L}_{pmml}^{M,V} + (1-p)\mathcal{L}_{pmml}^{M,I}. \tag{9}$$

Note that to further alleviate the modality imbalance problem, we adopt a weight p to balance the two losses, which is the same as the ratio of patch-mix. The effect of ratio p is that the more image information of one modality the mixed modality contains, the more similar it will be to the source modality.

3.4. Overall Optimization

Ultimately, we optimize the PMCM in an end-to-end manner with the final loss defined as follows:

$$\mathcal{L} = \mathcal{L}_{base} + \mathcal{L}_{c2c} + \mu\mathcal{L}_{pmml} \tag{10}$$

4. Experiments

4.1. Experimental Settings

4.1.1. Datasets

We evaluate our proposed framework on two VI-ReID datasets, SYSU-MM01 [17] and RegDB [34].

The SYSU-MM01 dataset contains 491 identities captured by 4 visible cameras and 2 infrared cameras both including indoor and outdoor environments. The training set contains 22258 visible images and 11909 infrared images involving 395 identities, while the testing set contains 96 identities with 3803 infrared images as query images. Following the protocols, we test it both in all-search mode and indoor-search mode for only single-shot.

The RegDB dataset contains 412 identities with 206 identities for training and 206 identities for testing, where each identity has 10 visible images and 10 infrared images from a pair of overlapping visible and infrared cameras. Following the protocols, we test it both in Visible2Thermal mode, where visible images as query and infrared images as the gallery, and Thermal2Visible mode similar to the former.

4.1.2. Evaluation metrics

For both datasets, we adopt the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) to evaluate the performance and take the average result of ten tests to report.

4.1.3. Implementation details

We implement our method with PyTorch and use ResNet50 pre-trained on ImageNet [35] as the backbone. All the input images are data augmented with a sequence of being resized to the size of $3 \times 384 \times 192$, random horizontal flipping, and random channel erasing [36]. The size of a mini-batch is set to 32, where we randomly sample 4 identities for each modality and 4 images for each identity. Besides, we adopt SGD as the optimizer with a weight decay of 5×10^{-4}, a momentum of 0.9, and a dynamic learning rate schedule, where the rate linearly increases from 0 to 0.1 in the first 10 epochs and decreases by 0.1 per 30 epochs after the 30th epoch. The total number of training epochs is set to 101 and the number of tests is 10. In terms of hyper-parameters, we set λ_1, λ_2 and λ_3 to 0.2, 0.2 and 1.0. Following [32], we set the margin of triplet loss to 0.3. The ratio p of patch-mix is set to 0.1 for SYSU-MM01 and 0.5 for RegDB. The parameter μ is gradually increased from 0 to 0.5 in the first 50 rounds and kept until the end.

We train our framework with one Tesla V100 GPU and the time cost of one epoch is increased by 54% compared to that of baseline. However, our inference time is exactly the same as the baseline, which is more important

Method	All-search								Indoor-Search							
	Single-Shot				Multi-Shot				Single-Shot				Multi-Shot			
	R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP
X-Modality [12]	49.92	89.79	95.96	50.73	-	-	-	-	-	-	-	-	-	-	-	-
MMN [11]	70.6	96.2	99.0	66.9	-	-	-	-	76.2	97.2	99.3	79.6	-	-	-	-
MPANet [31]	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
MCSL [8]	64.82	-	-	60.81	68.05	-	-	51.84	-	-	-	-	-	-	-	-
JCCL [22]	57.20	94.30	98.40	59.30	60.70	95.20	98.60	52.60	66.60	98.80	99.70	74.70	73.80	99.40	99.90	68.30
DCLNet [18]	70.79	-	-	65.18	-	-	-	-	73.51	-	-	76.80	-	-	-	-
FMCNet [23]	66.34	-	-	62.51	73.44	-	-	56.06	68.15	-	-	74.09	78.86	-	-	63.82
CMT [37]	71.88	96.45	98.87	68.57	80.23	97.91	99.53	63.13	76.98	97.68	99.64	79.91	84.87	99.41	99.97	74.11
MTL [19]	67.25	95.38	98.46	64.29	72.95	96.94	99.27	57.62	69.58	96.66	99.03	74.37	80.39	98.80	99.83	68.60
MTMFE [6]	69.47	96.42	99.11	66.41	73.74	97.48	99.59	59.78	71.72	97.19	98.97	76.38	81.49	99.18	99.79	70.94
G^2DA [20]	63.94	93.34	97.29	60.73	71.23	95.93	98.59	54.99	71.06	97.31	99.47	76.01	80.83	98.50	99.77	68.88
SIDA [3]	68.36	95.91	98.56	64.19	-	-	-	-	73.28	97.35	99.52	77.49	-	-	-	-
LDAEF [7]	66.61	-	-	62.86	75.24	-	-	56.70	70.90	-	-	75.78	81.84	-	-	69.42
PMT [38]	67.53	95.36	98.64	64.98	-	-	-	-	71.66	96.73	99.25	76.52	-	-	-	-
MRCN-P [39]	70.8	96.5	99.1	67.3	-	-	-	-	76.4	98.5	99.9	80.0	-	-	-	-
ProtoHPE [40]	71.92	96.19	97.98	70.59	-	-	-	-	77.81	98.64	99.59	81.31	-	-	-	-
CMM [24]	51.80	92.72	97.71	51.21	56.27	94.08	98.12	43.39	54.98	94.38	99.41	63.7	60.42	96.88	99.5	53.52
SMCL [15]	67.39	92.87	96.76	61.78	72.15	90.66	94.32	54.93	68.84	96.55	98.77	75.56	79.57	95.33	98.00	66.57
MID [14]	60.27	92.90	-	59.40	-	-	-	-	64.86	96.12	-	70.12	-	-	-	-
IMG [10]	61.31	91.31		57.20	69.79	95.12		51.01	67.20	96.06		72.41	78.14	97.69		65.51
PMCM(ours)	**75.54**	**97.49**	**99.30**	**71.16**	**82.52**	**99.00**	**99.78**	**65.88**	**81.52**	**98.99**	**99.71**	**84.33**	**90.06**	**99.80**	**99.97**	**79.45**

Table 1: Comparison of CMC and mAP performances with the SOTAs on SYSU-MM01. Particularly, methods in the last five lines adopt different mixup strategies to generate an intermediate modality for model learning.

in applications in the real-world scene.

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

We compare the proposed PMCM with several state-of-the-art methods for VI-ReID, including X-Modality [12], MMN [11], MPANet [31], SMCL [15], MCSL [8], JCCL [22], DCLNet [18], FMCNet [23], MID [14], CMT [37], SIDA [3], LDAEF [7], PMT [38], MRCN-P [39], ProtoHPE [40], MTL [19], MTMFE [6], G^2DA [20], IMG [10], and CMM [24].

The comparison results on SYSU-MM01 and RegDB are respectively shown in Table 1 and Table 2. We observe that our PMCM outperforms the existing SOTAs on all evaluation metrics by a large margin in SYSU-MM01. PMCM is superior to the second-best ProtoHPE in single-shot and all-search mode in SYSU-MM01 by 3.62% in Rank-1 accuracy and 0.57%

Method	Visible2Infrared		Infrared2Visible	
	Rank-1	mAP	Rank-1	mAP
X-Modality [12]	62.20	60.20	-	-
MMN [11]	91.6	84.1	87.5	80.5
MPANet [31]	82.8	80.7	83.7	80.9
MCSL [8]	93.83	87.55	91.55	85.25
JCCL [22]	78.8	69.4	77.9	69.4
DCLNet [18]	81.2	74.3	78.0	70.6
FMCNet [23]	89.12	84.43	88.38	83.86
CMT [37]	**95.17**	87.3	**91.97**	84.46
MTL [19]	89.91	85.64	88.34	84.06
MTMFE [6]	85.04	82.52	81.11	79.59
G^2DA [20]	73.95	65.49	69.67	61.98
SIDA [3]	81.73	75.07	79.71	72.60
LDAEF [7]	90.76	87.30	88.79	85.44
PMT [38]	84.83	76.55	84.16	75.13
MRCN-P [39]	95.1	89.2	92.6	86.5
ProtoHPE [40]	88.74	83.72	88.69	81.99
CMM [24]	59.81	60.86	-	-
SMCL [15]	83.93	79.83	83.05	78.57
MID [14]	87.45	84.85	84.29	81.41
IMG [10]	89.70	85.82	87.64	84.03
PMCM(ours)	93.09	**89.57**	91.44	**87.15**

Table 2: Comparison of CMC and mAP performances with the SOTAs on RegDB.

in mAP. Compared with the methods adopting mixup schemes, PMCM also shows the best, exceeding SMCL in single-shot and all-search mode in SYSU-MM01 by 8.15% in Rank-1 accuracy and 9.38% in mAP, demonstrating that our methods can encourage the model to learn the semantic correspondence between the two different modalities thus improving the overall performance.

Although some methods like MSCL, CMT, and MRCN-P surpass us in certain metrics in RegDB, their performance in SYSU-MM01 is significantly inferior to our PMCM. For instance, CMT exceeds our PMCM by 2.08% in Rank-1 accuracy in Visible2Infrared mode and 0.53% in Rank-1 accuracy in Infrared2Visible mode, but CMT is inferior to our PMCM by 3.66% in Rank-1 accuracy and 2.59% in mAP in single-shot and all-search mode in

	B	Part	PartAlign	C2C	PatchMix	PMML	Rank-1	mAP
1	✓	-	-	-	-	-	64.58	62.41
2	✓	✓	-	-	-	-	67.24	64.90
3	✓	✓	✓	-	-	-	69.52	65.37
4	✓	-	-	✓	-	-	68.63	64.43
5	✓	✓	✓	✓	-	-	70.66	66.06
6	✓	-	-	-	✓	-	67.92	63.77
7	✓	-	-	-	✓	✓	69.08	64.20
8	✓	✓	✓	✓	✓	-	73.76	69.88
9	✓	✓	✓	✓	✓	✓	75.54	71.16

Table 3: Ablation studies on the effectiveness of each component of the proposed PMCM in SYSU-MM01, where B denotes the baseline method.

SYSU-MM01. Compared with them, PMCM exhibits a better overall performance, providing evidence for effectively alleviating the modality imbalanced problem.

All the results above fully demonstrate the superiority and robustness of our PMCM, where more modality-invariant and discriminative features can be learned.

4.3. Algorithm Analysis

4.3.1. Ablation studies

To validate each component of PMCM, we conduct ablation experiments on SYSU-MM01 in the all-search and single-shot mode in an accumulation way. The experimental result is shown in Table 3 and numbered by row.

Effectiveness of the part-based baseline (Part). When exploring part features upon the global-based baseline, the performance is increased by 2.66% and 2.49% on Rank-1 and mAP respectively, showing its enhancement

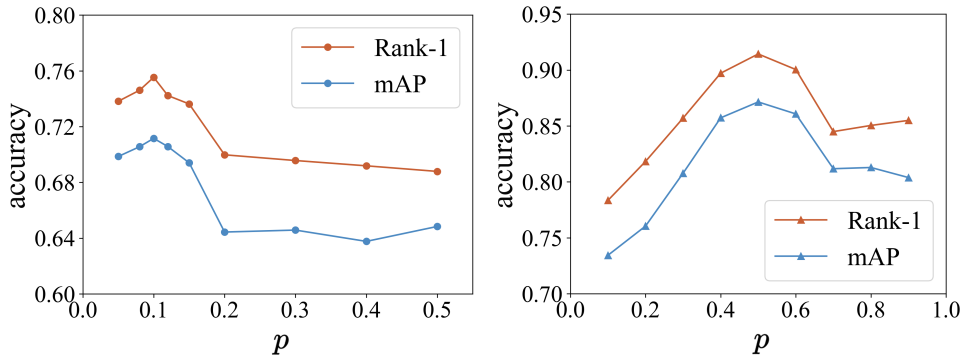
to the feature representation learning.

Effectiveness of the part alignment loss (PartAlign). The part alignment loss plays the role of regularization to the global and part feature predictions. Comparing row 3 with row 2 in the table, we observe that the improvements in Rank-1 accuracy and mAP are 2.32% and 0.47%, proving the part alignment loss better mines discriminative part features.

Effectiveness of the center-to-center loss (C2C). Compared with the baseline, C2C improves the Rank-1 accuracy and mAP by 4.05% and 2.02% and collaboration with "Part" and "PartAlign" further improves the two metrics by 2.03% and 1.63%, thanks to that C2C pulls close the centers of different modalities of the same identity to extract modality-invariant features.

Effectiveness of training with patch-mixed images (PatchMix). Adding PatchMix strategy to the baseline increases the Rank-1 accuracy and mAP by 3.34% and 1.36%, respectively, illustrating its effective. After introducing the modules above, the PatchMix strategy can still provide a significant improvement of the two metrics by 3.10% and 3.82%.

Effectiveness of patch-mixed cross-modality learning (PMML). As an auxiliary learning scheme, PMML shows improvement of the Rank-1 accuracy and mAP in two further experiments based on PatchMix, demonstrating its ability to enhance the effectiveness of PatchMix.



(a) SYSU-MM01 (b) RegDB

Figure 4: Influence of the different values of patch-mix ratio p, (a) experiments on SYSU-MM01 in all-search and single-shot mode, and (b) experiments on RegDB in Infrared2Visible mode.

4.3.2. Analysis of patch-mix ratio

The ratio p adjusts the proportion of IR patches and RGB patches in the patch-mixed images. When p is 0, the image is composed of only an IR image. As p increases, more RGB patches are contained. Based on this, we argue that the model will ultimately learn $(1 + p)$ times the RGB information and $(2 - p)$ times the IR modality information. When the modality imbalanced problem occurs, we can rebalance the RGB and IR information by adjusting the value of p. As shown in Fig. 4, we evaluate the ratio p on both SYSU-MM01 in all-search single-shot mode and RegDB in Infrared2Visible mode.

On SYSU-MM01, the samples of IR modality are much fewer than those of RGB modality. We observe that when the ratio p is set to 0.5 (the number of patches of two modalities is equal), a relatively low re-ID performance is obtained. When p decreases from 0.5 to 0.1 (more IR patches contained), the performance is continuously improved, and when the ratio is set to 0.1, the

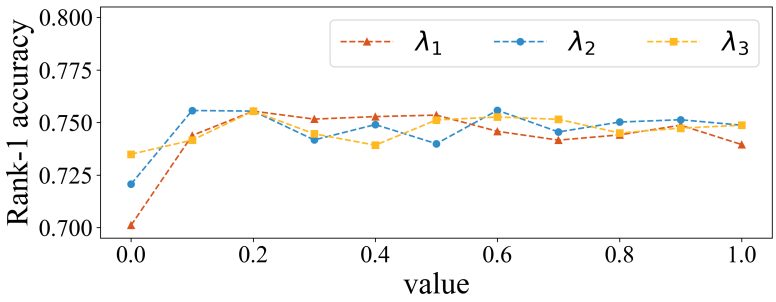


Figure 5: Analysis of the Rank-1 accuracy with parameters λ_1, λ_2 and λ_3. We keep the other parameters constant while testing the target one, and the table shows the insensitivity of our PMCM to these parameters.

best performance is achieved. This demonstrates that, when there is more IR information contained in the intermediate modality, a better balance of IR and RGB data is kept, and the model could learn the two modalities equally.

On RegDB, IR and RGB modalities have the same number of samples, which means RegDB is a data-balanced dataset. As shown in Fig. 4 (b), the model achieves best performance when p is set to 0.5. In addition, when the ratio is larger or smaller than 0.5, the balance between the two modalities is broken, thus leading to a performance decrease.

All pieces of evidence above prove that the proposed patch-mix scheme effectively alleviates the modality imbalance problem.

4.3.3. Analysis of parameters λ_1, λ_2 and λ_3

We have evaluated the parameters including λ_1, λ_2, and λ_3 by control variates in SYSU-MM01. The line chart of the rank-1 accuracy is shown in Fig. 5, which reveals that the fluctuation range of the metrics is approximately 3% within the range (0.1, 1.0) and if the value falls to 0, the accuracy

Method	Rank-1	Rank-10	mAP
CutMix [25]	69.48	95.07	64.73
RandomErasing [41]	69.71	95.31	63.95
Grayscale	70.94	95.95	65.84
PatchMix	**73.76**	**97.34**	**69.88**
Mixup [16] + PMML	73.26	96.95	69.39
RandomErasing [41] + PMML	73.74	97.02	70.00
CutMix [25] + PMML	73.90	97.16	70.31
PatchMix + PMML	**75.54**	**97.49**	**71.16**

Table 4: Comparison of different intermediate modality generation strategies on SYSU-MM01.

will significantly decrease. The experiments demonstrate that our method is insensitive to the value of three loss weights but each of them counts.

4.3.4. Analysis of different mixup strategies

In addition, to show the superiority of our method over other intermediate modality generation strategies, we compare images generated by grayscale, the standard mixup with PMML, CutMix [25], CutMix with PMML, RandomErasing [41] and RandomErasing with PMML. During implementation, we replace the PatchMix and PMML modules in our framework with the other generative methods and conduct experiments on them in the SYSU-MM01 dataset. The results are shown in Table 4. We observe that our method exceeds all of the strategies, which fully demonstrates the advantage of our patch-mix scheme over other existing methods. Moreover, our PMML scheme works quite effectively on the other intermediate modality generation strategies and produces different degrees of improvement in the two metrics.

method	Rank-1	Rank-10	Rank-20	mAP
Schedule1	71.89	96.38	98.63	67.14
Schedule2	73.91	97.19	98.94	69.57
ours	75.54	97.49	99.30	71.16

Table 5: Analysis of the update schedule. Schedule1 is constant at 0.5, and Schedule2 is a linearly increasing value.

Size	Rank-1	Rank-10	Rank-20	mAP
4×4	74.47	97.38	99.30	69.92
8×8	75.09	97.27	99.26	70.93
12×12	75.13	97.31	99.27	70.86
16×16	**75.54**	**97.49**	**99.30**	**71.16**
32×32	74.62	97.13	99.11	69.92

Table 6: Influence of different sizes of patch experimented on SYSU-MM01.

4.3.5. Analysis of the global-local balancing weight μ.

In order to speed up the convergence and reduce the error caused by the low-quality features extracted from part features in the early training stage, we design the special update schedule for the hyper-parameter, which linearly increases from 0 to 0.5 in the first half of the training phase and remains unchanged till the end. To validate the effectiveness of our schedule, we compare it with two other schedules. Schedule1 is constant at 0.5, and Schedule2 is a linearly increasing value represented by $current\_epoch/total\_epochs$. Based on the results shown in Table 5, our schedule achieves the best performance, which exceeds Schedule1 by 3.65% and Schedule2 by 1.63% in Rank-1 accuracy, which meets our expectations.

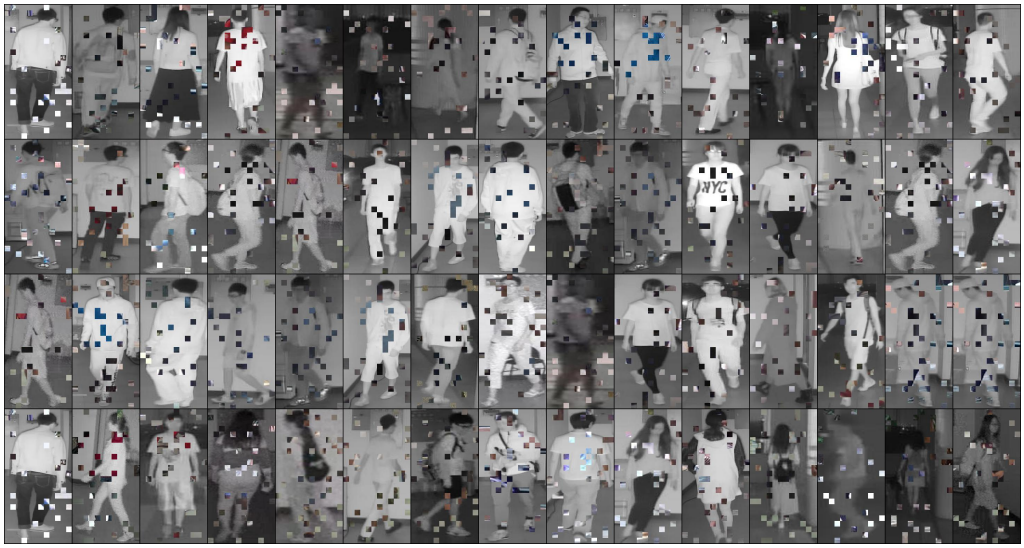


Figure 6: Visualization of PatchMixed images in SYSU-MM01 with the mixup ratio 0.1. It is obvious that all the generated images retain most of the infrared modality information and introduce a small amount of RGB modal information. Only very few of them have semantic misalignment issues, such as missing a piece on the face and replacing it with a piece from the surroundings.

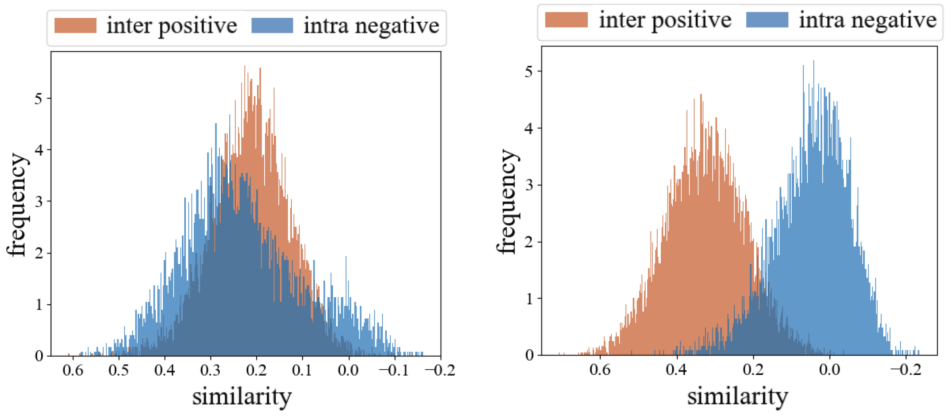
4.3.6. Analysis of different sizes of the patch

We conduct experiments to investigate the impact of different sizes of the image patches in SYSU-MM01. According to the results shown in Table6, either too large or too small sizes will slightly reduce the effect, which means our method is quite robust to the patch size. Finally, we choose the patch size of 16×16 for our method.

4.4. Visualization

4.4.1. Visualization of PatchMixed images

We manually checked 200 random PatchMixed images in SYSU-MM01 with the mixup ratio of 0.1 and part of the PatchMixed images are shown in



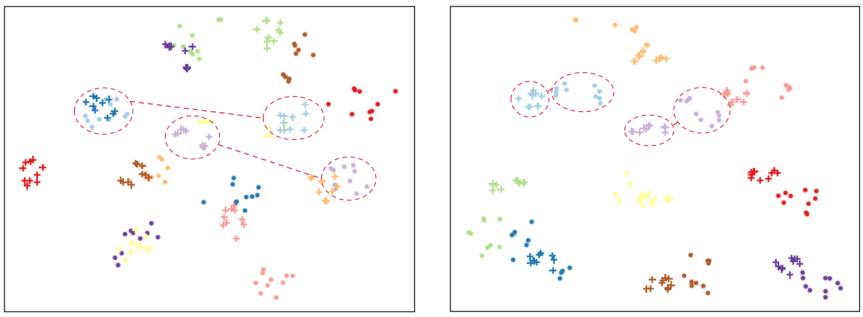
(a) Baseline (b) PMCM

Figure 7: Visualization of cosine similarity distribution of inter-modality positive samples and intra-modality negative samples with (a) baseline and (b) our PMCM on the test set of SYSU-MM01.

Fig. 6. As a result, we find that in most generated images, the mixed patches are harmonious with the patches around them, while less than 10% percent of generated mixed images have body part miss alignment between modalities. In addition, most misalignment is contour shifts within a body part. The case of containing two right hands (or two other body parts) has never happened. Although the edges of patches may not be always consistent, most of the time, the semantic meaning of a patch is consistent with its neighbors.

4.4.2. Cosine distance distribution

We visualize the cosine distance distribution of inter-modality positive samples and intra-modality negative samples in the test set of SYSU-MM01, as is shown in Fig. 7. In baseline (Fig. 7(a)), the two types of image pair share a similar distance distribution, which reveals that the baseline can



(a) Baseline (b) PMCM

Figure 8: Visualization of the feature embeddings distribution of (a) baseline and (b) PMCM via t-SNE in the SYSU-MM01 dataset, where dots and pluses in the same color denote features of the same identity in RGB modality and infrared modality.

hardly retrieve inter-modality positive images. Instead, our proposed PMCM (Fig. 7(b)) separates the inter-positive pairs and intra-negative pairs, showing that cross-modality images could have bigger similarity and be retrieved successfully. It fully demonstrates that our method can effectively reduce the intra-class modality discrepancy.

4.4.3. Feature distribution

To further explore the reason why PMCM is effective, we visualize the feature distribution of samples in the SYSU-MM01 dataset via t-SNE as shown in Fig. 8, where different colors denote different identities. For the baseline (Fig. 8(a)), feature embeddings of the same identity are far away, which indicates the baseline can hardly narrow the large modality gap. In comparison, our PMCM (Fig. 8(b)) discriminates and aggregates these feature embeddings of the same identity separately and clearly, demonstrating that our method can effectively improve representation learning and narrow

the cross-modality gap.

5. Conclusion

In this paper, we propose a Patch-Mixed Cross-Modality (PMCM) framework for VI-ReID. A patch-mixed modality is introduced to learn the semantic correspondence between visible and infrared images and alleviate the modality imbalance problem. A patch-mixed modality learning loss is adopted to take advantage of the third modality to reduce the modality gap between the two modalities. The patch-mixed modality and its learning strategy can be assumed as an add-on component and easily adopted in future work. To further reduce the inter- and intra-modality variance, we propose a part-alignment loss to constrain the consistency of part and global prediction distributions for more discriminative representation. Extensive experiment results have demonstrated the superior effectiveness of PMCM compared with other state-of-the-art methods. The code will be publicly available, which will enable future research. The limitation of our work is that some generated images may have body parts that miss alignment between modalities, which could lead to inaccurate representation learning. In our future work, we plan to explore solutions to improve alignment in patch-mixed images.

References

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. Hoi, Deep learning for person re-identification: A survey and outlook, TPAMI 44 (6) (2021) 2872–2893.

[2] A. Zahra, N. Perwaiz, M. Shahzad, M. M. Fraz, Person re-identification: A retrospective on domain specific open challenges and future trends, Pattern Recognition (2023) 109669.

[3] J. Gong, S. Zhao, K.-M. Lam, X. Gao, J. Shen, Spectrum-irrelevant fine-grained representation for visible–infrared person re-identification, Computer Vision and Image Understanding 232 (2023) 103703.

[4] Y. Gavini, A. Agarwal, B. Mehtre, Thermal to visual person re-identification using collaborative metric learning based on maximum margin matrix factorization, Pattern Recognition 134 (2023) 109069.

[5] P. K. Sarker, Q. Zhao, Enhanced visible–infrared person re-identification based on cross-attention multiscale residual vision transformer, Pattern Recognition 149 (2024) 110288.

[6] N. Huang, J. Liu, Y. Luo, Q. Zhang, J. Han, Exploring modality-shared appearance features and modality-invariant relation features for cross-modality person re-identification, Pattern Recognition 135 (2023) 109145.

[7] G. Zhang, Y. Zhang, H. Zhang, Y. Chen, Y. Zheng, Learning dual attention enhancement feature for visible-infrared person re-identification, Journal of Visual Communication and Image Representation (2024) 104076.

[8] Y. Ling, Z. Luo, Y. Lin, S. Li, A multi-constraint similarity learning with adaptive weighting for visible-thermal person re-identification., in: IJCAI, 2021, pp. 845–851.

[9] J. Zhu, H. Wu, Q. Zhao, H. Zeng, X. Zhu, J. Huang, C. Cai, Visible-infrared person re-identification using high utilization mismatch amending triplet loss, Image and Vision Computing 138 (2023) 104797.

[10] J. Lu, S. Zhang, M. Chen, X. Chen, K. Zhang, Cross-modality person re-identification based on intermediate modal generation, Optics and Lasers in Engineering 177 (2024) 108117.

[11] Y. Zhang, Y. Yan, Y. Lu, H. Wang, Towards a unified middle modality learning for visible-infrared person re-identification, in: ACM MultiMedia, 2021, pp. 788–796.

[12] D. Li, X. Wei, X. Hong, Y. Gong, Infrared-visible cross-modal person re-identification with an x modality, in: AAAI, Vol. 34, 2020, pp. 4610–4617.

[13] M. Ye, J. Shen, L. Shao, Visible-infrared person re-identification via homogeneous augmented tri-modal learning, TIFS 16 (2020) 728–739.

[14] Z. Huang, J. Liu, L. Li, K. Zheng, Z.-J. Zha, Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification, AAAI (2022).

[15] Z. Wei, X. Yang, N. Wang, X. Gao, Syncretic modality collaborative learning for visible infrared person re-identification, in: ICCV, 2021, pp. 225–234.

[16] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, arXiv preprint arXiv:1710.09412 (2017).

[17] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, J. Lai, Rgb-infrared cross-modality person re-identification, in: ICCV, 2017, pp. 5380–5389.

[18] H. Sun, J. Liu, Z. Zhang, C. Wang, Y. Qu, Y. Xie, L. Ma, Not all pixels are matched: Dense contrastive learning for cross-modality person re-identification, in: ACM MultiMedia, 2022, pp. 5333–5341.

[19] N. Huang, K. Liu, Y. Liu, Q. Zhang, J. Han, Cross-modality person re-identification via multi-task learning, Pattern Recognition 128 (2022) 108653.

[20] L. Wan, Z. Sun, Q. Jing, Y. Chen, L. Lu, Z. Li, G2da: Geometry-guided dual-alignment learning for rgb-infrared person re-identification, Pattern Recognition 135 (2023) 109150.

[21] S. Choi, S. Lee, Y. Kim, T. Kim, C. Kim, Hi-cmd: Hierarchical cross-

modality disentanglement for visible-infrared person re-identification, in: CVPR, 2020, pp. 10257–10266.

[22] Z. Zhao, B. Liu, Q. Chu, Y. Lu, N. Yu, Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification, in: AAAI, Vol. 35, 2021, pp. 3520–3528.

[23] Q. Zhang, C. Lai, J. Liu, N. Huang, J. Han, Fmcnet: Feature-level modality compensation for visible-infrared person re-identification, in: CVPR, 2022, pp. 7349–7358.

[24] Y. Ling, Z. Zhong, Z. Luo, P. Rota, S. Li, N. Sebe, Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification, in: ACM MultiMedia, 2020, pp. 889–897.

[25] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: ICCV, 2019, pp. 6023–6032.

[26] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: CVPR, 2019, pp. 9268–9277.

[27] C. Huang, Y. Li, C. C. Loy, X. Tang, Learning deep representation for imbalanced classification, in: CVPR, 2016, pp. 5375–5384.

[28] K. Cao, C. Wei, A. Gaidon, N. Arechiga, T. Ma, Learning imbalanced datasets with label-distribution-aware margin loss, NeurIPS 32 (2019).

[29] L. Shen, Z. Lin, Q. Huang, Relay backpropagation for effective learning of deep convolutional neural networks, in: ECCV, 2016, pp. 467–482.

[30] J. Liu, Y. Sun, F. Zhu, H. Pei, Y. Yang, W. Li, Learning memory-augmented unidirectional metrics for cross-modality person re-identification, in: CVPR, 2022, pp. 19366–19375.

[31] Q. Wu, P. Dai, J. Chen, C.-W. Lin, Y. Wu, F. Huang, B. Zhong, R. Ji, Discover cross-modality nuances for visible-infrared person re-identification, in: CVPR, 2021, pp. 4330–4339.

[32] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv preprint arXiv:1703.07737 (2017).

[33] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: ECCV, 2018, pp. 480–496.

[34] D. T. Nguyen, H. G. Hong, K. W. Kim, K. R. Park, Person recognition system based on a combination of body images from visible light and thermal cameras, Sensors 17 (3) (2017) 605.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: CVPR, Ieee, 2009, pp. 248–255.

[36] M. Ye, W. Ruan, B. Du, M. Z. Shou, Channel augmented joint learning for visible-infrared recognition, in: ICCV, 2021, pp. 13567–13576.

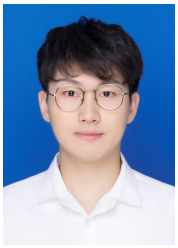
[37] K. Jiang, T. Zhang, X. Liu, B. Qian, Y. Zhang, F. Wu, Cross-modality transformer for visible-infrared person re-identification, in: ECCV, Springer, 2022, pp. 480–496.

[38] H. Lu, X. Zou, P. Zhang, Learning progressive modality-shared transformers for effective visible-infrared person re-identification, in: AAAI, Vol. 37, 2023, pp. 1835–1843.

[39] Y. Zhang, Y. Yan, J. Li, H. Wang, Mrcn: A novel modality restitution and compensation network for visible-infrared person re-identification, CVPR (2023).

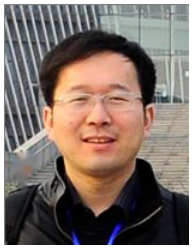
[40] G. Zhang, Y. Zhang, Z. Tan, Protohpe: Prototype-guided high-frequency patch enhancement for visible-infrared person re-identification, in: ACM MultiMedia, 2023, pp. 944–954.

[41] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: AAAI, Vol. 34, 2020, pp. 13001–13008.



Zhihao Qian received the B.E. degree from Wuhan University, China, in 2022. He is currently a master's student at Wuhan University, China. His research interests are person re-ID and unsupervised learning.

Yutian Lin received the B.E. degree from Zhejiang University, China, in 2016, and the Ph.D. degree from the Center for Artificial Intelligence, University of Technology Sydney, Australia, in 2019. She is currently an associate professor in Wuhan University, China. Her research interests include person re-ID and related applications, unsupervised learning and self-supervised learning.



Bo Du (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the State Key Lab of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2010. He is currently a Professor with National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, Wuhan University, Wuhan, China. His major research interests include pattern recognition, hyperspectral image processing, machine learning, and signal processing.