Channel Augmented Joint Learning for Visible-Infrared Recognition

Mang Ye[1], Weijian Ruan[2], Bo Du[1]*, Mike Zheng Shou[3]
[1] National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence,
Hubei Key Laboratory of Multimedia and Network Communication Engineering,
School of Computer Science, Wuhan University, China
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
[3] National University of Singapore, Singapore

https://github.com/mangye16/Cross-Modal-Re-ID-baseline

Abstract

This paper introduces a powerful channel augmented joint learning strategy for the visible-infrared recognition problem. For data augmentation, most existing methods directly adopt the standard operations designed for single-modality visible images, and thus do not fully consider the imagery properties in visible to infrared matching. Our basic idea is to homogenously generate color-irrelevant images by randomly exchanging the color channels. It can be seamlessly integrated into existing augmentation operations without modifying the network, consistently improving the robustness against color variations. Incorporated with a random erasing strategy, it further greatly enriches the diversity by simulating random occlusions. For cross-modality metric learning, we design an enhanced channel-mixed learning strategy to simultaneously handle the intra- and cross-modality variations with squared difference for stronger discriminability. Besides, a channel-augmented joint learning strategy is further developed to explicitly optimize the outputs of augmented images. Extensive experiments with insightful analysis on two visible-infrared recognition tasks show that the proposed strategies consistently improve the accuracy. Without auxiliary information, it improves the state-of-the-art Rank-1/mAP by 14.59%/13.00% on the large-scale SYSU-MM01 dataset.

1. Introduction

Identity recognition (person re-identification [22], face recognition [11, 19]) systems have achieved significant success recently. However, most research efforts have been paid on the single-modality visible domain. In many night-time surveillance and low-light environments, near(far)-infrared cameras are applied to capture target appearance

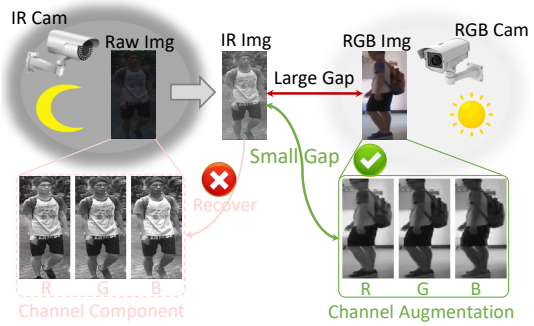


Figure 1. Motivation of channel augmentation. Directly recovering the three-channel RGB image from a single-channel infrared images is ill-posed. Instead, we propose to directly optimize the relation between the infrared images and the R, G and B channels of the visible images.

[4, 37]. This raises important cross-modality visible-infrared recognition problems, *e.g.*, visible-infrared person re-identification (VI-ReID) [41] and NIR-VIS face recognition [53]. The cross-modality matching is usually formulated by learning modality shared or invariant features.

Matching the infrared imagery to visible-spectrum images is a significant challenge due to the large modality gap and unknown environmental factors [32, 54] (*e.g.*, different viewpoints, occlusions, background clutter, etc.), leading to large intra- and cross-modality variations. To eliminate the color discrepancy, cross-modality image generation with Generative Adversarial Networks (GANs) is a popular approach [36, 56] bridging the gap at the image level. However, the image generation process usually needs additional computational cost and suffers from unavoidable noise [21]. Another approach is to directly employ grayscale images to perform the cross-modality matching [9, 42], where the color information is assumed to be irrelevant. While this approach does eliminate the color discrepancy, it also loses discriminative information in color channels.

This paper presents a channel exchangeable augmenta-

*Corresponding Author: Bo Du

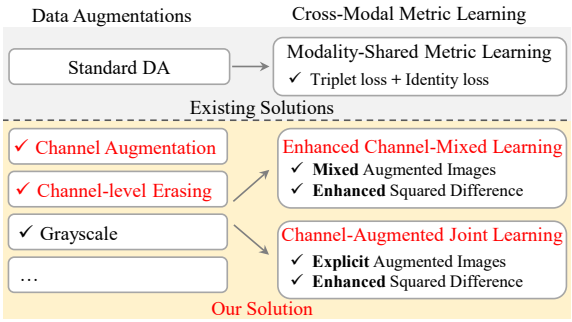


Figure 2. Illustration of the major contributions. Two important steps: data augmentation (DA) and cross-modal metric learning.

tion (CA) to narrow the gap at the input image level, while keeping informative color information. An illustration of the main idea is shown in Fig. 1. A straightforward solution to eliminate the modality discrepancy is to recover the original three color channels. However, transforming a single-channel infrared image to a three-channel visible image is a challenging problem with unavoidable noise [52]. Alternatively, we propose to directly learn the relation between each R, G and B channel of the visible images and the single-channel infrared images. This serves as a channel augmentation operation for the visible-to-infrared learning process to reinforce the robustness against color variations. We further present a random erasing (CRE) technique for occlusion simulations. Incorporated with the channel augmentation, our strategy performs the erasing in the channel level for better diversity. Besides, we also include a grayscale transformation for augmentation, reducing the color effect. These augmentation operations greatly enlarge the training set, bringing in better generalizability.

For cross-modality metric learning, we first present an enhanced channel-mixed learning scheme. Different from the widely-used bidirectional triplet metric learning [7, 24, 49], we directly optimize the feature embeddings in a **mixed** batch with the same identity classifier and metric for original visible, infrared and channel augmented modalities. Specifically, we design a weighted regularized triplet loss with an **enhanced squared difference** for the cross-modality metric learning, simultaneously handling the intra- and cross-modality variations. Our design has two primary benefits: 1) It fully considers all the possible triplet relations in the augmented image set. 2) The squared difference approximates the large margin metric learning principle [40] to improve the discriminability. We also develop a channel-augmented joint learning strategy to **explicitly** optimize the channel augmented images for training. The basic idea is to treat the channel augmented visible images as an additional modality, formulating a tri-modality joint learning framework. It slightly increases the computational burden at each training step, but consistently improves the testing accuracy without additional cost. Our main contributions (summarized in Fig. 2) are:

- We present a novel channel exchangeable augmentation for visible-infrared recognition. It can be seamlessly integrated into existing augmentations operations without modifying the network structure or changing the learning strategy.
- We design an enhanced channel-mixed learning scheme to simultaneously handle the intra- and cross-modality variations. With a joint learning strategy, it explicitly optimizes the channel augmented images.
- We evaluate on both visible-infrared person re-identification and face recognition, achieving significant accuracy gains under various settings.

2. Related Work

Visible-Infrared Person Re-IDentification (VI-ReID) is a cross-modality person recognition problem, which aims at matching daytime visible and night-time infrared images [2, 48, 16, 57]. Wu *et al.* [42] started the first attempt by introducing a zero-padding one-stream network, where they directly utilized grayscale images for training and testing. To simultaneously handle the intra- and cross-modality variations, a bi-directional dual-constrained framework was presented in [49]. Additionally, Dai *et al.* [7] also proposed to jointly discriminate the identity and modality in an adversarial training framework. The ID-discriminative factors and ID-excluded factors are disentangled in [5]. To fully utilize the relations across two modalities, a dual-attentive aggregation learning method was designed in [50]. With the advancement of GANs, a dual-level discrepancy modeling method [36] generates the cross-modality images, eliminating discrepancy at pixel-level. An improved version with an alignment constraint was developed in [33]. But the image generation process introduces unavoidable noise.

Visible Infrared Face Recognition (VI-FR) (a.k.a heterogeneous FR) [9, 14, 27, 46] is closely related to VI-ReID, with the goal of matching face images across two modalities. Learning modality-related metrics or dictionaries with hand-craft features [39, 18] is the popular approach. For deep learning, most methods focus on learning sharable multi-modality features [46], cross-modality matching models [28] or disentangled representations [45]. The lightCNN model [44] is served as a powerful baseline for VI-FR [53]. Recently, a Pose Aligned Cross-spectral Hallucination (PACH) approach [9] was designed to disentangle the independent factors at multiple stages.

Data Augmentation has been widely applied in many different computer vision tasks [20, 30, 47, 60]. It enlarges the training set through various translations, such as cropping, rotation, flipping, adding noises, mix-Up [55], etc. The random erasing [59] is applied in many fine-grained recognition problems for better generalizability. Recently, several auto-augmentation techniques [6] are developed for various applications. However, most of these methods are designed for single-modality visible images/videos.

3. Proposed Channel Augmentation

The learning target of cross-modality visible-infrared matching is that the image features of the same class under different modalities are invariant. We denote the original cross-modality training set as $\mathcal{T} = \{\mathcal{T}^v, \mathcal{T}^r\}$. In particular, $\mathcal{T}^v = \{\mathbf{x}_i^v | i = 1, 2, \cdots, N^v\}$ represents the visible training set with N^v visible images, where each image $\mathbf{x}_i^v = \{x_i^R, x_i^G, x_i^B\}$ is composed of three channels, *i.e.*, R, G and B. $\mathcal{T}^r = \{\mathbf{x}_i^r | i = 1, 2, \cdots, N^r\}$ represents N^r infrared training images, where each element $\mathbf{x}_i^r$ is an infrared image with a single over-saturated grayscale channel. The cross-modality visible-infrared matching aims at learning a feature extraction network $f^v \in \mathcal{F}$ for the visible modality and $f^r \in \mathcal{F}$ for the infrared modality[1]. The learning objective is to optimize the relation between the extracted features $f^v(x_i^{R,G,B})$ of visible images in three-channel RGB space and the features $f^r(x_j^r)$ of near-infrared images in single channel space, denoted by

$$\mathcal{L} = \sum \ell(f^v(x_i^{R,G,B}), f^r(x_j^r), y_i, y_j), \tag{1}$$

where y_i and y_j are the annotated training labels for each image. $\ell(\cdot)$ is an objective function for optimizing the relations, which can be identity loss [58], triplet loss [15] or their variants [51]. Note that our proposed augmentation strategies in this section can be seamlessly integrated into various baseline models without modifying the learning strategy or network architectures.

3.1. Random Channel Exchangeable Augmentation

The basic motivation behind the Channel exchangeable Augmentation (CA) is that the three-channel color visible images contain rich appearance information, and the color information is beneficial for the visible-infrared matching. However, directly recovering a three-channel visible image from a single-channel infrared image is quite challenging. Instead, we explicitly learn to match the infrared images and color channels of the visible images. Specifically, we introduce a channel augmentation strategy by mining the relation between each individual channel (R, G or B) and the single-channel infrared images. The main idea is to randomly select one channel (R, G or B) to replace the other channels, which generates a new training image by concentrating on one single channel. This is formulated as

$$\begin{aligned} \tilde{x}_i^{v,R} &= (x_i^R, x_i^R, x_i^R) \\ \tilde{x}_i^{v,G} &= (x_i^G, x_i^G, x_i^G) \\ \tilde{x}_i^{v,B} &= (x_i^B, x_i^B, x_i^B). \end{aligned} \tag{2}$$

Some visualization results of the augmentation are shown in Fig. 3. We observe that the channel augmented images share similar visual appearance with the infrared images

[1]The two modalities may share the same feature network, as in [7, 36].

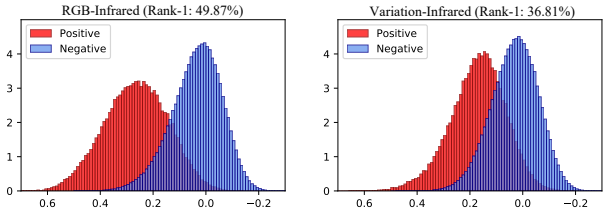
Figure 3. Illustration of the channel exchangeable augmentation in visible-infrared person re-identification.

while maintaining the original texture structure of the visible images. With the channel augmented images, the learning objective of visible-infrared matching becomes

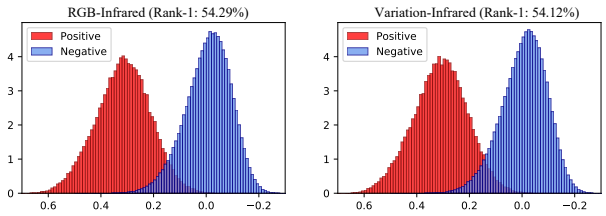
$$\mathcal{L} = \sum \ell(f^v(\tilde{x}_i^v), f^r(x_j^r), y_i, y_j), \tag{3}$$

where $\tilde{x}_i^v$ represents a randomly channel augmented visible image or its original three-channel RGB image. The implementation of the channel exchangeable augmentation training is straightforward, and introduces a minimal computation overhead. It can be seamlessly integrated with other basic data augmentation operations (*random flipping*, *random resizing* and *random cropping*). We use a single data loader to perform the random channel augmentation, which does not increase the size of mini-batch input. Right after the general image transform functions, we add the random channel augmentation function. It first chooses a random integer number from $[0, 1, 2, 3]$. This value determines whether the original RGB image is kept or the random channel augmentation is performed as in Eq. (2). This strategy does not introduce additional I/O communication and only slightly increases the computational cost in the transformation process. The testing protocol is the same as standard setting, where we do not contain any additional augmentations for fair comparison with existing methods.

What is CA doing? The channel augmentation can be understood as a homogenous generation of three-channel visible images by decomposing the color channels. This strategy encourages the model f to learn the explicit relation between each color channel of the visible images and the single-channel infrared images. To demonstrate the learned robustness against color variations, we visualize the pairwise positive similarity scores (belonging to the same identity) and negative similarity score distributions of (belonging to the different identities) in Fig. 4. We train two baseline models using AGW [51], with or without the channel augmentation. Both models are trained under the same setting on the visible-infrared person re-identification dataset, SYSU-MM01 [42]. We evaluate both RGB-to-



(a) Without Channel Augmentation



(b) With Channel Augmentation

Figure 4. Robustness of the Channel Augmentation (CA). Visualization of the cross-modality matching in the testing set under different training settings. x-axis denotes the cosine similarity score, y-axis represents a normalized value for each quantized similarity bin of 10000 randomly selected positive/negative matching pairs. With channel augmentation, we observe that the separations of both RGB-Infrared and Variation-Infrared matching are better than those of without channel augmentation.

infrared (*left figures*) and variation-to-infrared (*right figures*) matching results on the testing set. For *variation-to-infrared matching*, we random apply the grayscale transformation or channel augmentation to the RGB images, simulating the color variations. Specifically, the similarity scores are calculated for 10000 randomly selected positive/negative pairs, where x-axis denotes the cosine similarity score and y-axis represents a normalized value for each quantized similarity bin. We also report the rank-1 matching accuracy under different query settings in Fig. 4.

From this experiment, we draw three interesting conclusions: *1) Channel augmentation enhances the invariance for the positive matching pairs*, *i.e.*, the pairwise positive similarity scores of "*w CA*" are generally much larger than those of "*w/o CA*" ("red bins" in Fig. 4). It demonstrates that the variance is also reduced when channel augmentation is applied. This means that the model trained with CA is more stable in terms of input color variations. *2) Channel augmentation also introduces a larger difference for the negative matching pairs*, *i.e.*, the pairwise negative similarities are also slightly decreased ("blue bins" in Fig. 4). The main reason is that the random color variations bring in larger appearance change for the unmatched negative sample pairs, introducing the larger variance. *3) The proposed channel augmentation greatly improves the representation robustness against color variations*. With channel augmentation, the proposed model achieves much better separation for both RGB-infrared and variation-infrared matching. In terms of the rank-1 accuracy, we observe that the variation-to-infrared matching is extremely bad (upper right in Fig. 4) if without the channel augmentation, while we achieve consistent satisfying performance under the same setting. This further demonstrates the robustness against color variations.

Figure 5. Illustration of the channel-level random erasing (CRE) in visible-infrared face recognition. Note that the color of the channel-level erased image is just for illustration.

3.2. Channel-Level Random Erasing

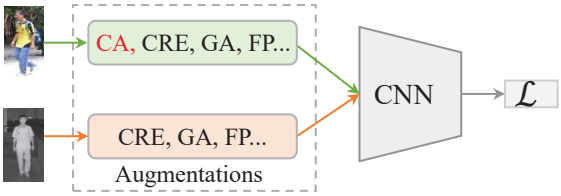
This section presents a channel-level random erasing scheme for visible-infrared matching. Random erasing [59] has been widely evaluated on various vision tasks to improve the generalizability on testing task [25]. Given a predefined erasing probability, its basic idea is to randomly select a rectangular region I_e within a training image, and replaces its pixel values with random values for all three channels, simulating the uncertain occlusions. In short, it is an *image-level* random patch erasing.

Incorporating with the channel augmentation, we design a *channel-level* random erasing (CRE) strategy to enrich the variety of training samples. Specifically, assume that the size of a three-channel visible training image is $W \times H \times 3$. We randomly select the erasing area of a rectangular region S_e, the size of which erasing area is bounded with a specific ratio. Together with the channel augmentation, we randomly select the erasing area for different channels (R, G and B). Within the selected erasing region S_e^* for each channel, each pixel in S_e^* is assigned to a specific pre-defined value α^*, where * represents the corresponding channel index. Empirically, we select the mean values of the R, G and B channels obtained from the large-scale ImageNet [8] as the erasing value for each channel. Generally, the formulation of the channel-level random erasing is defined as follows

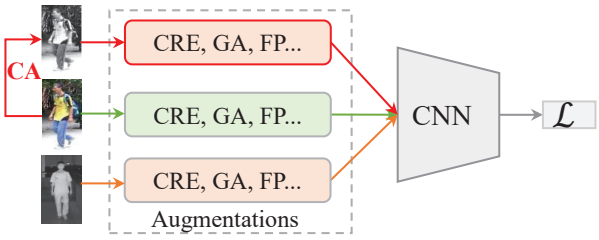
$$\tilde{x}_i^{v,*}(m,n) = \begin{cases} \alpha^*, & (m,n) \in S_e^* \\ \tilde{x}_i^{v,*}(m,n), & otherwise \end{cases} \quad (4)$$

where m and n represent the coordinate position of a pixel. α^* is calculated by the average value for each channel. For a single-channel infrared image, we simply transform it to three copied single-channel image in the channel-level random erasing process. Some example images with the channel-level augmentation for the visible to near-infrared face recognition task are shown in Fig. 5.

The proposed channel-level random erasing has two primary advantages: 1) It further enriches the augmentations at



(a) Enhanced Channel-Mixed Learning in § 4.1



(b) Channel-Augmented Joint Learning in § 4.2

Figure 6. Comparison of the channel-augmented joint learning in § 4.2 and the channel-mixed learning strategy in § 4.1. CA: channel augmentation, CRE: channel-level random erasing, GA: grayscale augmentation, FP: horizontal-flip.

the channel level, providing richer supervision for the cross-modality feature representation learning. Together with the channel augmentation, the erased images greatly enlarge the training sample set. 2) The erased images also improve the robustness against image noise, e.g., partial occlusions, imperfect detections. The two augmentations are both easy to implement and consistently improve the performance.

Other Augmentations. Considering that the grayscale transformation (GA) also serves as simple baseline to transform the three-channel RGB images to single-channel grayscale image [52], we propose to incorporate with a random grayscale transformation as an augmentation supplement operation, enhancing the robustness against color variations for visible-infrared recognition. Besides, we also evaluate the a random horizontal flip (FP) operation to address the viewpoint variation for person recognition.

4. Cross-Modality Metric Learning

This section presents two cross-modality metric learning strategies, enhanced channel-mixed learning in § 4.1 and channel-augmented joint learning in § 4.2. An illustration is shown in Fig. 6.

4.1. Enhanced Channel-Mixed Learning

Baseline. General cross-modality matching models usually apply bi-directional triplet variants to guide the cross-modality feature learning [23, 38, 49], optimizing the relative distance between the cross-modal positive and negative pairs. However, this strategy does not effectively address the intra-modality variations. To simultaneously handle the intra- and cross-modality variations, this paper firstly adopts a channel-mixed learning strategy, constructing a batch that contains images from different modalities. It directly optimizes the relations without considering the modality difference. Specifically, it is a combination of the identity classification loss ($\mathcal{L}_{id}$) and weighted regularization triplet loss ($\mathcal{L}_{wrt}$) [51]. The identity loss $\mathcal{L}_{id}$ treats images of the same identity across two modalities as the same class. It is represented by

$$\mathcal{L}_{id} = -\frac{1}{N}\sum_{i=1}^{N}\log(p(y_i|f(x_i);\theta^0)), \tag{5}$$

where θ^0 represents the shared identity classifiers for both channel augmented visible images and infrared images under different data augmentation operations. $f(x_i)$ is a general function for extracting the features of images from different modalities. It can be different for two modalities.

The weighted regularization triplet loss aims at optimizing the relative distance between all the positive and negative pairs, from both intra-modality and cross-modality relations. Similar to [15], a `softplus` function is adopted for the optimization, which is represented by

$$\mathcal{L}_{wrt} = \frac{1}{N}\sum_{i=1}^{N}\log(1+\exp(\textstyle\sum_{ij} w_{ij}^p d_{ij}^p - \sum_{ik} w_{ik}^n d_{ik}^n)), \tag{6}$$

$$w_{ij}^p = \frac{\exp\left(d_{ij}^p\right)}{\sum_{d_{ij}^p\in\mathcal{P}_i}\exp(d_{ij}^p)}, w_{ik}^n = \frac{\exp\left(-d_{ik}^n\right)}{\sum_{d_{ik}^n\in\mathcal{N}_i}\exp(-d_{ik}^n)},$$

where (i, j, k) represents a triplet within each training batch for each anchor sample x_i. Note that j and k can be from either the same modality or different modalities in the channel-mixed learning strategy. For anchor x_i, $\mathcal{P}_i$ is the corresponding positive set and $\mathcal{N}_i$ is the negative set. d_{ij}^p/d_{ik}^n represents the pairwise distance of a positive/negative sample pair. d_{ij} is the Euclidean distance between two samples, represented by $d_{ij} = \|f(x_i) - f(x_j)\|_2$. The weighting strategy with softmax function greatly increases the contribution of hard samples with larger (smaller) distance for positives (negatives). Unlike the hard triplet mining [15, 43], our proposed strategy fully utilizes all the sampled triplets within each batch by adaptively considering their contributions. Meanwhile, it directly optimizes the relative distance between the positive and negative pairs for both intra- and inter modality learning, resulting in stronger robustness against variations.

Enhanced Squared Difference. A widely-used technique to measure the pair distance difference is ℓ_1 norm [29, 31, 34], as done Eq. 6. This paper introduces an enhanced squared difference. The basic idea is to optimize the squared difference rather than the ℓ_1 norm difference, which is represented by

$$\mathcal{L}_{sq} = \frac{1}{N}\sum_{i=1}^{N}\log(1+\exp(\phi[\underbrace{\textstyle\sum_{ij} w_{ij}^p d_{ij}^p - \sum_{ik} w_{ik}^n d_{ik}^n}_{\mu_i}])),$$

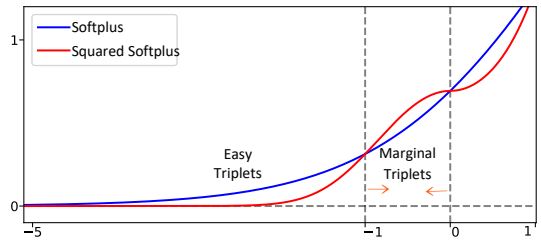


Figure 7. Illustration of the enhanced `squared` difference for `softplus`. We observe that it increases the contribution of the marginal triplets ($\sum_{ij} w_{ij}^p d_{ij}^p - \sum_{ik} w_{ik}^n d_{ik}^n) \in [-1, 0]$, while reducing the effect of easy triplets ($\sum_{ij} w_{ij}^p d_{ij}^p - \sum_{ik} w_{ik}^n d_{ik}^n) < -1$. This approximates the idea of large margin metric learning [40], but our design does not require additional margin parameter.

$$\phi[\mu_i] = \begin{cases} \mu_i^2, & \mu_i > 0, \\ -\mu_i^2, & \mu_i < 0. \end{cases} \tag{7}$$

The above formula replaces the original $\sum_{ij} w_{ij}^p d_{ij}^p - \sum_{ik} w_{ik}^n d_{ik}^n$ with a squared variant. To demonstrate the effect of the modification, we plot the curves of original `softplus` function and the enhanced `squared` version in Fig. 7. Geometrically, it has three primary benefits:

- It increases the contribution of the *marginal hard triplets* ($\mu_i \in [-1, 0]$) in the overall learning objective. These *marginal triplets* already satisfy the constraint $\sum_{ij} w_{ij}^p d_{ij}^p < \sum_{ik} w_{ik}^n d_{ik}^n$, but their discriminability is still limited since the contribution in the overall model is decreased. The `square` operation slightly enlarges the difference for stronger discriminability, but does not require additional hyper-parameter.
- It also reduces the effect of *easy triplets* ($\mu_i < -1$). These triplets have already satisfied the constraints, and the difference is relative large, *i.e.*, $\sum_{ij} w_{ij}^p d_{ij}^p + 1 < \sum_{ik} w_{ik}^n d_{ik}^n$. The contribution thus should be decreased in the overall learning objective.
- It slightly decreases the influence of *primitive triplets* ($\mu_i \in [-1, 0]$). These triplets are far from the optimized status and usually have a very large value for the overall loss calculation. Decreasing their contribution would enforce that the optimization concentrates more on the marginal hard triplet.

The enhanced channel-mixed learning directly optimizes the features with the same identity classifier and distance metric for both (channel augmented) infrared and visible modalities. With a squared difference, this results in stronger robustness against intra-modality variations and consistently improves the cross-matching performance.

4.2. Channel-Augmented Joint Learning

This section designs a channel-augmented joint learning strategy to fully utilize the channel augmented visible images. Specifically, we explicitly treat the channel augmented visible images as an auxiliary modality. Together with the images from original visible and infrared modalities, our proposed strategy formulates a tri-modality joint learning framework, in Fig. 6 (b). Similar to the enhanced channel-mixed learning in § 4.1, we use a combination of identity loss and enhanced weighted regularized triplet loss as the training objective, denoted by

$$\mathcal{L} = \mathcal{L}_{id} + \mathcal{L}_{sq}. \tag{8}$$

The major difference is that the channel augmented images act as an additional modality, but they share the same identity classifier as the infrared and visible images. This strategy can enforce the model to focus on learning modality-invariant feature representations. As an alternative, we also try to apply separate classifiers for different modalities [10], but it does not bring consistent improvement. Meanwhile, other advanced cross-modality matching models might also be integrated to improve the feature learning process.

Discussion. The channel-augmented joint learning strategy fully utilizes the channel augmented images without modifying the network structures. With the same number of visible and infrared input images, it requires large memory in the training process but it keeps the same as standard settings in the testing phase. Another benefit is that this strategy formulates a large batch size, providing more informative hard samples for the cross-modality feature learning. Extensive experiments have validated the consistent improvements under various settings.

5. Experimental Results

5.1. Experiments on VI-ReID

We first evaluate our propose model on the visible-infrared person re-identification (VI-ReID) task, including two public datasets (SYSU-MM01 [42] and RegDB [26]). Following the settings in [42, 49], rank-k matching accuracy and mean Average Precision (mAP) are used as evaluation metrics. In addition, we also report the mean Inverse Negative Penalty (mINP) metric proposed in [51], which measures the cost for finding all the correct matches.

SYSU-MM01 [42] is the largest VI-ReID dataset, and it is collected at SYSU campus under both indoor and outdoor environments during the daytime and night-time. It is captured by 4 RGB and 2 near-infrared cameras. The training set contains 22,258 visible and 11,909 near-infrared images of 395 identities. This datasets contains both *all-search* and *indoor-search* testing mode, where the former mode is more challenging. Detailed dataset settings can be found in [42].

RegDB [26] dataset is captured by one visible and one far-infrared (thermal) camera. It comprises 412 person identities in total, and each identity has 10 visible and 10 far-infrared images. Following existing VI-ReID settings, we randomly sample 206 identities for training and the remaining 206 identities are used for testing. The training/testing split procedure is repeated ten times [33, 36, 51],

Table 1. Evaluation of each augmentation component on two VI-ReID datasets. "B" represents the baseline results with default data augmentations. "CA" denotes the random channel exchangeable augmentation, "CRE" means the random channel erasing, and "GA" is the random grayscale augmentation. "FP" is the horizontal flip operation augmentation. "J" indicates the channel-augmented joint learning strategy (§ 4.2) with enhanced squared triplet loss in Eq. (7). Rank at r accuracy(%), mAP (%) and mINP (%) are reported.

	Augmentations				SYSU-MM01 *(All Search)*				SYSU-MM01 *(Indoor Search)*				RegDB			
B	CA	CRE	GA	FP	$r=1$	$r=5$	mAP	mINP	$r=1$	$r=5$	mAP	mINP	$r=1$	$r=5$	mAP	mINP
Base					49.87	77.30	49.76	37.63	55.64	84.24	64.01	60.61	75.87	85.58	68.88	52.55
	✓				54.29	81.44	51.56	36.91	60.45	87.01	66.97	62.47	77.27	85.86	68.45	54.42
		✓			59.38	84.19	58.26	45.76	63.14	88.39	69.79	65.74	78.16	88.06	73.54	60.51
	✓	✓			65.28	88.83	61.97	47.43	71.33	92.17	76.04	71.91	82.27	89.69	73.45	61.17
	✓	✓	✓		65.91	89.77	62.98	48.68	72.28	92.42	77.38	73.38	82.39	90.34	73.57	61.76
	✓	✓		✓	66.11	89.44	63.03	48.80	72.56	92.74	77.08	73.09	82.59	90.06	74.31	61.82
	✓	✓	✓	✓	67.02	90.44	64.43	50.05	73.18	92.52	77.47	73.56	83.51	91.49	77.06	62.74
J	✓	✓	✓	✓	69.88	90.98	66.89	53.61	76.26	94.42	80.37	76.79	85.03	92.48	79.14	65.33

and the overall average accuracy is reported for comparison.

Implementation Details. This work is supported by Huawei MindSpore [1]. Following AGW [51], we adopt a non-local module enhanced two-stream network with ResNet50 [13] as the feature extraction backbone. The network parameters are initialized with the ImageNet pre-trained weights. For the baseline data augmentation, we apply the random cropping with zero-padded images (288×144) for the training images. For infrared images, three copied channels are fed into the network. For the optimization, Stochastic Gradient Descent (SGD) optimizer is adopted for training. The initial learning rate is set to 0.1, and decayed by 0.1 and 0.01 at 20, 50 epochs. We apply a warm-up strategy [25] in the first 10 epochs. The total number of training epoch is 100. At each training step, we randomly sample 8 identities, of which 4 visible and 4 infrared images are selected to formulate a batch.

5.2. Ablation Study

Effect of Each Augmentation Operation. We firstly evaluate the effect of each augmentation operation on two VI-ReID datasets. The results are shown in Table 1.

1) Effectiveness of CA: We observe significant improvement in accuracy under various settings when applying our designed random channel exchangeable augmentation. This augmentation also greatly improves the robustness against color variations. *2) Effectiveness of CRE*: When integrating the random channel erasing for data augmentation, the accuracy increases for all three metrics. This simulates the random uncertain occlusion in the training process, which greatly improves the generalizability on the testing set. When incorporating with the CA, the performance is further dramatically reinforced. *3) Effectiveness of GA*: We also observe a slight improvement for the random grayscale augmentation operation, which simulates a single-channel transformed dataset for three channel visible images. When applying the *flip augmentation (FP)*, the robustness against different viewpoints is enhanced. The experiments demonstrate that all these augmentation components contribute consistently to the overall model accuracy gain.

Comparison of Different Learning Schemes. As shown in Table 1, with channel augmented joint training (J), the rank-1 accuracy is further improved from 67.02% to 69.88% on the large-scale SYSU-MM01 dataset . This demonstrates that the proposed method can benefit from explicit channel-augmented multi-modality joint learning.

Table 2. Evaluation of `squared` difference (Eq. 7) on the SYSU-MM01. "HardTri" represents the online hard triplet mining. "M" denotes the enhanced channel-mixed learning in § 4.1 and "J" means the channel augmented joint learning in § 4.2.

Strategy	*All Search*			*Indoor Search*		
	$r=1$	mAP	mINP	$r=1$	mAP	mINP
HardTri	66.41	64.35	51.05	72.43	77.15	73.03
MS [35]	68.54	65.53	52.12	73.23	77.92	73.68
M (w/o Sq)	67.02	64.43	50.05	73.18	77.47	73.56
M	69.16	64.97	50.27	73.84	78.24	74.33
J (w/o Sq)	67.68	64.62	51.87	73.92	78.40	74.74
J	69.88	66.89	53.61	76.26	80.37	76.79

Table 3. Applicability of our proposed channel augmentation with state-of-the-art methods on the large-scale SYSU-MM01 dataset.

Strategy	*All Search*			*Indoor Search*		
	$r=1$	mAP	mINP	$r=1$	mAP	mINP
DDAG [50]	54.75	53.02	39.62	61.02	67.98	62.61
+ Ours	62.51	57.76	43.82	68.84	73.82	69.24
	↑7.76	↑4.74	↑4.20	↑7.82	↑5.84	↑6.63

This part evaluates the enhanced squared difference in Eq. (7) on the SYSU-MM01 dataset in Table 2. Compared to the widely-used online hard triplet mining [15] and multi-similarity loss [35], our strategy generally performs better under various settings. Compared to the original ℓ_1 norm difference, the enhanced `squared` difference achieves in consistent improvement for all the metrics. These results validate our analysis in § 4.1, *i.e.*, our method enhances the discriminability of hard triplets. Comparison with "M" further verifies the advantage of joint learning.

Applicability with Other Methods. We also evaluate our proposed augmentations incorporated into DDAG [50]. We use the default settings with the authors' released code. The results in Table 3 demonstrate that the performance can be significantly improved under various metrics. It would be further improved by fine-tuning the hyper-parameters.

5.3. Comparison with State-of-the-Arts

We also compare with the state-of-the-art VI-ReID methods published in the last two years, as shown in Tables 4 and 5. They demonstrate that our proposed model significantly outperforms existing solutions under various settings. There are three major advantages of our method: *1)*

Table 4. Comparison with the state-of-the-arts on SYSU-MM01 [42]. Rank at r accuracy (%), mAP (%) and mINP (%) are reported.

Settings		*All Search*					*Indoor Search*				
Method	Venue	$r=1$	$r=10$	$r=20$	mAP	mINP	$r=1$	$r=10$	$r=20$	mAP	mINP
Zero-Pad [42]	ICCV-17	14.80	54.12	71.33	15.95	-	20.58	68.38	85.79	26.92	-
HCML [48]	AAAI-18	14.32	53.16	69.17	16.16	-	24.52	73.25	86.73	30.08	-
eBDTR [49]	TIFS-19	27.82	67.34	81.34	28.42	-	32.46	77.42	89.62	42.46	-
HSME [12]	AAAI-19	20.68	32.74	77.95	23.12	-	-	-	-	-	-
D^{2}RL [36]	CVPR-19	28.9	70.6	82.4	29.2	-	-	-	-	-	-
AlignGAN [33]	ICCV-19	42.4	85.0	93.7	40.7	-	45.9	87.6	94.4	54.3	-
X-Modal [21]	AAAI-20	49.9	89.8	96.0	50.7	-	-	-	-	-	-
Hi-CMD [5]	CVPR-20	34.9	77.6	-	35.9	-	-	-	-	-	-
cm-SSFT [24]†	CVPR-20	47.7	-	-	54.1	-	-	-	-	-	-
AGW [51]	arXiv-20	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23
DDAG [50]	ECCV-20	54.75	90.39	95.81	53.02	39.62	61.02	94.06	98.41	67.98	62.61
HAT [52]	TIFS-20	55.29	92.14	97.36	53.89	-	62.10	95.75	99.20	69.37	-
Ours	-	**69.88**	**95.71**	**98.46**	**66.89**	**53.61**	**76.26**	**97.88**	**99.49**	**80.37**	**76.79**

† cm-SSFT [24] reported a higher matching accuracy by using all gallery samples as auxiliary information, which is infeasible in many applications.

Table 5. Comparison with the state-of-the-arts on RegDB [26]. Rank at r accuracy (%), mAP (%) and mINP (%) are reported.

Settings		*Visible to Infrared*					*Infrared to Visible*				
Method	Venue	$r=1$	$r=10$	$r=20$	mAP	mINP	$r=1$	$r=10$	$r=20$	mAP	mINP
Zero-Pad [42]	ICCV-17	17.75	34.21	44.35	18.90	-	16.63	34.68	44.25	17.82	-
HCML [48]	AAAI-18	24.44	47.53	56.78	20.08	-	21.70	45.02	55.58	22.24	-
eBDTR [49]	TIFS-19	34.62	58.96	68.72	33.46	-	34.21	58.74	68.64	32.49	-
HSME [12]	AAAI-19	50.85	73.36	81.66	47.00	-	50.15	72.40	81.07	46.16	-
D^{2}RL [36]	CVPR-19	43.4	66.1	76.3	44.1	-	-	-	-	-	-
AlignGAN [33]	ICCV-19	57.9	-	-	53.6	-	56.3	-	-	53.4	-
XModal [21]	AAAI-20	62.21	83.13	91.72	60.18	-	-	-	-	-	-
Hi-CMD [5]	CVPR-20	70.93	86.39	-	66.04	-	-	-	-	-	-
cm-SSFT [24]	CVPR-20	72.3	-	-	72.9	-	71.0	-	-	71.7	-
AGW [51]	arXiv-20	70.05	86.21	91.55	66.37	50.19	70.49	87.21	91.84	65.90	51.24
DDAG [50]	ECCV-20	69.34	86.19	91.49	63.46	49.24	68.06	85.15	90.31	61.80	48.62
HAT [52]	TIFS-20	71.83	87.16	92.16	67.56	-	70.02	86.45	91.61	66.30	-
Ours	-	**85.03**	**95.49**	**97.54**	**79.14**	**65.33**	**84.75**	**95.33**	**97.51**	**77.82**	**61.56**

Table 6. Evaluation of our proposed channel augmentation on two visible-infrared face recognition datasets. Rank at 1 accuracy (%) and false acceptance rate (F: %) are reported.

	Oulu [3]			*BUAA* [17]		
Strategy	$r=1$	F:1%	F:0.1%	$r=1$	F:1%	F:0.1%
IDR [14]	94.3	73.4	46.2	94.3	93.4	84.7
ADFL [46]	95.5	83.0	60.7	95.2	95.3	88.0
VSA[53]	99.9	96.8	82.3	98.0	98.2	92.5
PACH [9]	100	97.9	88.2	98.6	98.0	93.5
B	100	97.9	87.0	98.0	97.7	93.7
B + Ours	100	98.9	91.7	98.3	98.2	94.5

We do not need additional image generation [5, 33, 36] or adversarial training process [7, 24]. This property makes our proposed model more applicable for practical model deployment. We assume that the performance would be further improved if advanced cross-modality matching models were introduced, which could be seamlessly integrated with our proposed channel augmented joint learning strategy. 2) Our proposed solution does not contain any manually defined hyper-parameters. *3) The learned representation is robust against different cross-modality matching settings.* Notably, on the large-scale SYSU-MM01 dataset, we improve the Rank-1 accuracy and the mAP score by 14.59% and 13.00%, respectively.

5.4. Visible-Infrared Face Recognition

To demonstrate the generalizability, we also evaluate it on the visible-infrared face recognition task. We use Oulu-CASIA NIR-VIS [3] and BUAA-VisNir face databases [17]. We adopt LightCNN-29 [44] proposed in [9] as the baseline. Other training details and hyper-parameters are exactly the same. The rank-1 accuracy, VR@FAR=1%, and VR@FAR=0.1% are reported. We apply our channel augmentation together with the random erasing strategies to the baseline. The results are shown in Table 6.

We observe that our proposed augmentations also consistently improve the visible-infrared face recognition performance. This experiment further verifies that our method can be a general tool for visible-infrared matching tasks.

6. Conclusion

This paper presents a novel random channel augmentation method for visible infrared matching. It can be seamlessly integrated into different baseline methods without modifying the network structures, significantly improving the cross-modality recognition. We also present a channel-augmented joint learning strategy with enhanced squared difference to further reinforce the discriminability. We may further investigate the augmentation properties in other visible-infrared applications.

Acknowledgement. This work is partially supported by CAAI-Huawei MindSpore Open Fund, the National Natural Science Foundation of China under Grants (62176188, 61822113), the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170. Mike Shou is only supported by National Research Foundation Singapore (NRF-NRFF13-2021-0008). The numerical calculations in this paper had been supported by the super-computing system in the Supercomputing Center of Wuhan University.

References

[1] Mindspore, https://www.mindspore.cn/, 2020. 7

[2] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, volume 6, page 7, 2017. 2

[3] Jie Chen, Dong Yi, Jimei Yang, Guoying Zhao, Stan Z Li, and Matti Pietikainen. Learning mappings for face synthesis from near infrared to visual light images. In *CVPR*, pages 156–163, 2009. 8

[4] Ziang Cheng, Yinqiang Zheng, Shaodi You, and Imari Sato. Non-local intrinsic decomposition with near-infrared priors. In *ICCV*, pages 2521–2530, 2019. 1

[5] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *CVPR*, pages 10257–10266, 2020. 2, 8

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 2

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

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 4

[9] Boyan Duan, Chaoyou Fu, Yi Li, Xingguang Song, and Ran He. Cross-spectral face hallucination via disentangling independent factors. In *CVPR*, pages 7930–7938, 2020. 1, 2, 8

[10] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE TIP*, 29:579–590, 2020. 6

[11] Guodong Guo and Na Zhang. A survey on deep learning based face recognition. *CVIU*, 189:102805, 2019. 1

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

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[14] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Learning invariant deep representation for nir-vis face recognition. In *AAAI*, pages 2000–2006, 2017. 2, 8

[15] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3, 5, 7

[16] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *CVPR*, pages 7183–7192, 2019. 2

[17] D Huang, J Sun, and Y Wang. The buaa-visnir face database instructions. *School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001*, 2012. 8

[18] De-An Huang and Yu-Chiang Frank Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *ICCV*, pages 2496–2503, 2013. 2

[19] Brendan F Klare and Anil K Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE TPAMI*, 35(6):1410–1422, 2012. 1

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012. 2

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

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

[23] Hong Liu, Rongrong Ji, Yongjian Wu, Feiyue Huang, and Baochang Zhang. Cross-modality binary code learning via fusion similarity hashing. In *CVPR*, pages 6345–6353, 2017. 5

[24] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*, pages 13379–13389, 2020. 2, 8

[25] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *arXiv preprint arXiv:1906.08332*, 2019. 4, 7

[26] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. 6, 8

[27] Chunlei Peng, Nannan Wang, Jie Li, and Xinbo Gao. Re-ranking high-dimensional deep local representation for nir-vis face recognition. *IEEE TIP*, 2019. 2

[28] M Saquib Sarfraz and Rainer Stiefelhagen. Deep perceptual mapping for cross-modal face recognition. *International Journal of Computer Vision*, 122(3):426–438, 2017. 2

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 5

[30] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019. 2

[31] 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 6398–6407, 2020. 5

[32] Guangcong Wang, Jian-Huang Lai, Wenqi Liang, and Guangrun Wang. Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification. In *CVPR*, pages 10568–10577, 2020. 1

[33] Guanan 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 3623–3632, 2019. 2, 6, 8

[34] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, pages 5265–5274, 2018. 5

[35] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*, pages 5022–5030, 2019. 7

[36] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shinichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*, pages 618–626, 2019. 1, 2, 3, 6, 8

[37] Zheng Wang, Zhixiang Wang, Yinqiang Zheng, Yang Wu, Wenjun Zeng, and Shin'ichi Satoh. Beyond intra-modality: A survey of heterogeneous person re-identification. *arXiv preprint arXiv:1905.10048*, 2019. 1

[38] Jiwei Wei, Xing Xu, Yang Yang, Yanli Ji, Zheng Wang, and Heng Tao Shen. Universal weighting metric learning for cross-modal matching. In *CVPR*, pages 13005–13014, 2020. 5

[39] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE TCYB*, 47(2):449–460, 2017. 2

[40] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10(2), 2009. 2, 6

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

[42] Ancong Wu, Wei-shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianghuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017. 1, 2, 3, 6, 8

[43] Dongming Wu, Mang Ye, Gaojie Lin, Xin Gao, and Jianbing Shen. Person re-identification by context-aware part attention and multi-head collaborative learning. *IEEE Transactions on Information Forensics and Security (TIFS)*, 2021. 5

[44] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE TIFS*, 13(11):2884–2896, 2018. 2, 8

[45] Xiang Wu, Huaibo Huang, Vishal M Patel, Ran He, and Zhenan Sun. Disentangled variational representation for heterogeneous face recognition. In *AAAI*, pages 9005–9012, 2019. 2

[46] Xiang Wu, Lingxiao Song, Ran He, and Tieniu Tan. Coupled deep learning for heterogeneous face recognition. In *AAAI*, pages 1679–1686, 2018. 2, 8

[47] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019. 2

[48] 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, 8

[49] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C. Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE TIFS*, 15:407–419, 2020. 2, 5, 6, 8

[50] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *ECCV*, 2020. 2, 7, 8

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

[52] Mang Ye, Jianbing Shen, and Ling Shao. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE TIFS*, 2020. 2, 5, 8

[53] Aijing Yu, Haoxue Wu, Huaibo Huang, Zhen Lei, and Ran He. Lamp-hq: A large-scale multi-pose high-quality database for nir-vis face recognition. *arXiv preprint arXiv:1912.07809*, 2019. 1, 2, 8

[54] Shijie Yu, Shihua Li, Dapeng Chen, Rui Zhao, Junjie Yan, and Yu Qiao. Cocas: A large-scale clothes changing person dataset for re-identification. In *CVPR*, pages 3400–3409, 2020. 1

[55] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2

[56] He Zhang, Benjamin S Riggan, Shuowen Hu, Nathaniel J Short, and Vishal M Patel. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *IJCV*, 127(6-7):845–862, 2019. 1

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

[58] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 3

[59] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 2, 4

[60] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camstyle: A novel data augmentation method for person re-identification. *IEEE TIP*, 28(3):1176–1190, 2019. 2