Joint Color-irrelevant Consistency Learning and Identity-aware Modality Adaptation for Visible-infrared Cross Modality Person Re-identification

Zhiwei Zhao[1,2], Bin Liu[1,2,✉] , Qi Chu[1,2], Yan Lu[1,2], Nenghai Yu[1,2]

[1] School of Information Science and Technology, University of Science and Technology of China, Hefei, China
[2] Key Laboratory of Electromagnetic Space Information, Chinese Academy of Science, Hefei, China
zwzhao98@mail.ustc.edu.cn, {flowice, qchu}@ustc.edu.cn, luyan17@mail.ustc.edu.cn, ynh@ustc.edu.cn

Abstract

Visible-infrared cross modality person re-identification (VI-ReID) is a core but challenging technology in the 24-hours intelligent surveillance system. How to eliminate the large modality gap lies in the heart of VI-ReID. Conventional methods mainly focus on directly aligning the heterogeneous modalities into the same space. However, due to the unbalanced color information between the visible and infrared images, the features of visible images tend to overfit the clothing color information, which would be harmful to the modality alignment. Besides, these methods mainly align the heterogeneous feature distributions in dataset-level while ignoring the valuable identity information, which may cause the feature misalignment of some identities and weaken the discrimination of features. To tackle above problems, we propose a novel approach for VI-ReID. It learns the color-irrelevant features through the color-irrelevant consistency learning (CICL) and aligns the identity-level feature distributions by the identity-aware modality adaptation (IAMA). The CICL and IAMA are integrated into a joint learning framework and can promote each other. Extensive experiments on two popular datasets SYSU-MM01 and RegDB demonstrate the superiority and effectiveness of our approach against the state-of-the-art methods.

Introduction

Person re-identification (ReID) aims to match the images of a given person across multiple non-overlapping cameras (Zheng, Yang, and Hauptmann 2016). It is a core technology in large-scale intelligent video surveillance analysis. Many works focus on addressing the problem of person ReID under a single RGB modality, i.e., the visible-visible person images matching. In order to alleviate the difficulties caused by viewpoint changes, pose variations and occlusion, these methods mainly focus on robust feature learning (Liu and Zhang 2019; Qian et al. 2018; Zheng et al. 2019a; Wei et al. 2017; Sun et al. 2018) and effective metric learning (Varior, Haloi, and Wang 2016; Hermans, Beyer, and Leibe 2017; Chen et al. 2017), which have achieved excellent performance in public academic datasets. However, the assumption of the single RGB modality inevitably weakens the application scope of the person ReID. In a practical 24-hour video surveillance system, the visible cameras are hard to be deployed in night time or dark environment. In contrast, the infrared (IR) cameras using the infrared light to capture persons work well and play a complementary role with the visible cameras. Thus, it is necessary to study the visible-infrared cross modality person re-identification (**VI-ReID**) task. Given a visible (or infrared) image of a specific person, VI-ReID aims to retrieve all the infrared (or visible) images belonging to the same identity.

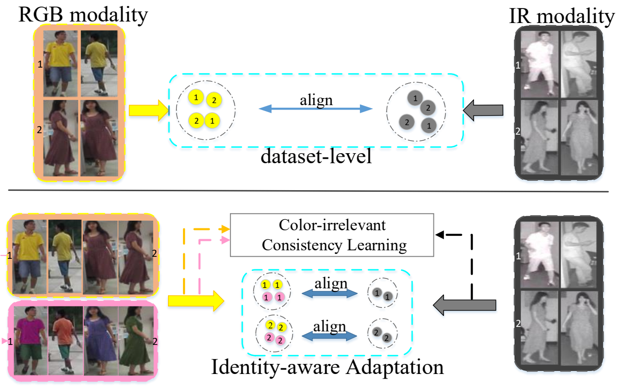

RGB modality
IR modality
align
dataset-level
Color-irrelevant Consistency Learning
align
align
Identity-aware Adaptation

Figure 1: Top: Conventional methods directly align the RGB modality and IR modality in dataset-level. Bottom: Our proposed method incorporates the color-irrelevant feature learning and identity-level modality adaptation together to effectively bridge the modality gap.

Compared with the traditional person ReID that only existing intra-modality discrepancy, VI-ReID encounters the additional large modality gap originating from the heterogeneous imaging processes of different cameras. For visible images in RGB modality, the rich clothing colors are very discriminative cues to distinguish different identities (Zheng et al. 2019b; Yang, Wu, and Zheng 2020), while the color cues are unavailable for infrared images in IR modality. To decrease the large modality gap, conventional methods focus on aligning RGB and IR modality in image space or feature space. Some GAN-based works (Wang et al. 2019b; Kniaz et al. 2018; Wang et al. 2020) attempt to leverage image translation to overcome pixel-wise discrepancy. However, recovering corresponding visible images from infrared images is an ill-posed problem, since a person in IR modality may wear diverse color clothes in RGB modality. Other typical methods (Wu et al. 2017; Ye et al. 2018a; Dai et al. 2018; Lu et al. 2020) leverage adversarial learning or metric learning to directly align the cross modality features into a same space. However, it is hard to directly align RGB modality and IR modality, because the features of RGB modalitiy tend to overfit the clothing color information, which would be harmful to the modality alignment. Besides, as shown in the top of Fig. 1, these methods mainly focus on aligning the feature distributions of entire RGB and IR images in dataset-level, while ignoring the valuable built-in identity information, which may result in the features of some identities are misaligned and harm the discrimination of the features.

To address the aforementioned problems, we propose a novel approach (termed CIMA) for VI-ReID. As shown in the bottom of Figure 1, the proposed method integrates the color-irrelevant feature learning and identity-level modality adaptation into a unified learning framework to effectively decrease the modality gap.

Specifically, our approach involves two novel components, i.e., the color-irrelevant consistency learning (CICL) and the identity-aware modality adaptation (IAMA). We propose the CICL to facilitate the color-irrelevant feature learning and minimize the negative effect of the clothing color overfitting to modality alignment. It applies the random clothing color transformation to visible images and further imposes both the intra and inter modality color-irrelevant consistency constraints to the visible images which share the same identity but with different color clothes. With the color-irrelevant consistency learning, the color-irrelevant features with high discriminability could be learned and the modality gap can be preliminarily decreased. Concurrently, the IAMA is proposed to adapt the identity-level features distributions into a same space. It fully exploits the built-in valuable identity information and performs modality adaptation within every identity, which effectively tackles the problem of feature misalignment of some identities and retains the discriminability of the learned features. The CICL and IAMA are integrated into a joint learning framework and can promote each other. We conduct extensive experiments on two popular datasets SYSU-MM01 (Wu et al. 2017) and RegDB (Nguyen et al. 2017) to demonstrate the superiority and effectiveness of our proposed approach. Our contributions are summarized as follows:

- We propose a novel color-irrelevant consistency learning method for VI-ReID task, which aims to learn color-irrelevant features with high discriminability by imposing color-irrelevant consistency constraints to random clothing color transformation.
- We present to adapt the feature distributions of RGB and IR modality with an identity-aware manner, which could retain the discriminability of features and tackle the problem of feature misalignment of some identities.
- We combine the color-irrelevant consistency learning and identity-aware modality adaptation in a unified framework, which significantly surpasses the state-of-the-art methods on VI-ReID datasets: SYSU-MM01 and RegDB.

Related Works

Visible-infrared Person Re-identification

The previous works in VI-ReID focused on directly aligning the RGB and IR modality in feature-space or image-space. Wu *et al.* (Wu et al. 2017) were the first to investigate VI-ReID and contributed a large-scale dataset SYSU-MM01. Ye *et al.* (Ye et al. 2018a,b) presented dual-constrained top-ranking loss to decrease the cross-modality variations. Dai *et al.* (Dai et al. 2018) proposed cross-modality generative adversarial network (cmGAN) to align the feature distributions of heterogeneous modalities into a same space by adversarial learning. Wang *et al.* (Wang et al. 2019a) introduced an alignment Adversarial Network (AlignGAN) by exploiting pixel alignment and feature alignment jointly. In (Wang et al. 2020), they further proposed to generate cross-modality paired-images and perform both set-level and instance-level alignments. Li *et al.* (Li et al. 2020) proposed to learn an auxiliary X modality and performed the X-Infrared-Visible (XIV) learning, while Ye *et al.* (Ye, Shen, and Shao 2020) presented the grayscale augmented Tri-Modal learning method (HAT) to better decrease the modality gap. However, few above works have attempted to explicitly minimize negative effect of the clothing color overfitting to modality alignment. Some recent works attempt to explore the influence of both modality-shared and modality-specific features. Choi *et al.* (Choi et al. 2020) (Hi-CMD) presented to disentangle the modality-shared and specific features and only used modality-shared features for retrieval. Lu *et al.* (Lu et al. 2020) proposed a cross-modality shared-specific feature transfer algorithm (cm-SSFT) to utilize both the modality shared and specific information, which achieves the state-of-the art performance. However, the Hi-CMD did not involve explicit modality adaptation, while cm-SSFT performed dataset-level modality adaptation.

Consistency Learning

Consistency learning has been widely exploited in semi-supervised learning (Sajjadi, Javanmardi, and Tasdizen 2016; Xie et al. 2019; Laine and Aila 2017; Berthelot et al. 2019), which leverages the idea that model should make consistent prediction for the inputs that undergoing data transformations but leave class semantics unaffected. Sajjadi *et al.* (Sajjadi, Javanmardi, and Tasdizen 2016) proposed to apply the consistency regularizatons by introducing the invariance to stochastic transformations and perturbations. Xie *et al.* (Xie et al. 2019) investigated the noise injection by quality data augmentation methods to improve consistency training. Our approach also shares the analogous idea by enforcing the color-irrelevant consistency constraints to the introduced random clothing color transformation.

Methodology

The overview of the proposed method is illustrated in Figure 2. The input images including the original visible images, the clothing color transformed visible images and infrared images are first fed into the two-stream network to extract the features. Then the color-irrelevant consistency

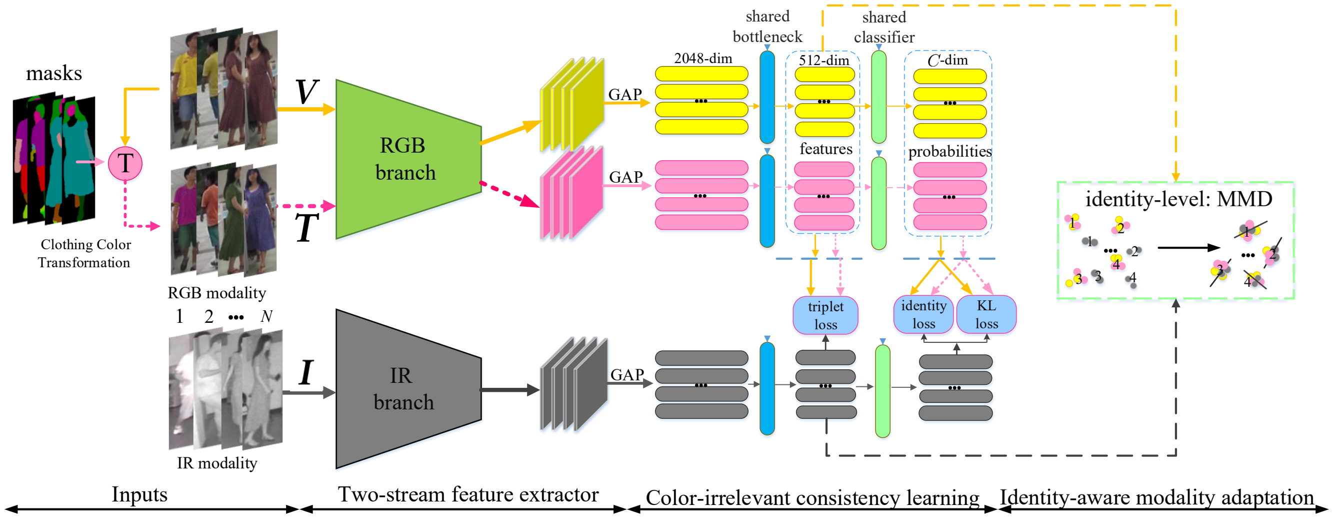


Figure 2: The framework of our proposed approach. The entire framework involves novel key components: the color-irrelevant consistency learning (CICL) and the identity-aware modality adaptation (IAMA). The CICL aims to facilitate the discriminative color-irrelevant feature learning, while the IAMA is presented to adapt the identity-level cross modality feature distributions simultaneously. The two components are jointly optimized in an end-to-end manner and can benefit each other.

learning (CICL) imposes both the intra-modality and inter-modality color-irrelevant consistency constraints to visible images which share the same identity but with different color clothes, which aims to facilitate the discriminative color-irrelevant feature learning. Concurrently, the identity-aware modality adaptation (IAMA) is presented to adapt the cross modality feature distributions into a modality-shared space in identity-level. The two components are integrated into a unified framework and can facilitate each other.

Two-stream Baseline

In this section, we introduce the adopted two-stream baseline. As shown in Fig. 2, given input original visible image set $\boldsymbol{V} = \{V_i\}_{i=1}^N$ and infrared image set $\boldsymbol{I} = \{I_i\}_{i=1}^N$, where $V_i \in R^{3\times H\times W}$, $I_i \in R^{3\times H\times W}$ denote the i^{th} corresponding image of $\boldsymbol{V}$ and $\boldsymbol{I}$, respectively. N is the number of visible or infrared images in a batch, 3 is the number of image channel. H and W are image height and width, respectively. We sample $\boldsymbol{V}$ and $\boldsymbol{I}$ from training set to ensure that the corresponding image pair, i.e., V_i and I_i for each $i \in \{1, \cdots, N\}$, belongs to the same identity. Then we utilize two-stream network (Ye et al. 2018b) consisting of modality-specific RGB and IR branches to extract 2048-dimensional features, and the backbone parameters of RGB and IR branch are not shared. GAP denotes global average pooling layer. The 2048-dimensional feature vectors are further fed into a modality-shared bottleneck consisting of sequential fully-connected (FC) layer, batchNorm (BN) layer to obtain the 512-dimensional feature vectors. We denote the extracted feature vector set of $\boldsymbol{V}$ and $\boldsymbol{I}$ as $\boldsymbol{F^V}$ and $\boldsymbol{F^I}$, respectively, where $\boldsymbol{F^V} = [\boldsymbol{f}_1^{\boldsymbol{V}}, \boldsymbol{f}_2^{\boldsymbol{V}}, \cdots, \boldsymbol{f}_i^{\boldsymbol{V}}, \cdots, \boldsymbol{f}_N^{\boldsymbol{V}}]$, $\boldsymbol{F^I} = [\boldsymbol{f}_1^{\boldsymbol{I}}, \boldsymbol{f}_2^{\boldsymbol{I}}, \cdots, \boldsymbol{f}_i^{\boldsymbol{I}} \cdots, \boldsymbol{f}_N^{\boldsymbol{I}}]$. $\boldsymbol{f}_i^{\boldsymbol{V}}$ and $\boldsymbol{f}_i^{\boldsymbol{I}}$ denote the 512-dimensional feature vector of V_i and I_i, respectively. To make sure the extracted features for both modalities are discriminative, we append a LeakyReLU layer and a modality-shared C-dimensional FC layer as classifier and employ the identity loss for identity-discriminative feature learning.

$$\mathcal{L}_{identity}^{\boldsymbol{VI}} = \mathbb{E}_{i,\boldsymbol{M}}[-\log(p(y_i^{\boldsymbol{M}}|\boldsymbol{f}_i^{\boldsymbol{M}})] \tag{1}$$

where $p(y_i^{\boldsymbol{M}}|\boldsymbol{f}_i^{\boldsymbol{M}})$ is the predicted probability of belonging to the ground-truth class $y_i^{\boldsymbol{M}}$ for the i^{th} input image of $\boldsymbol{M}$, and $\boldsymbol{M} \in \{\boldsymbol{V}, \boldsymbol{I}\}$. Besides, we also add the hard triplet loss (Wang et al. 2019b) to enhance the discriminative feature learning, which aims to decrease the intra-identity distance and enlarge the inter-identity distance. The hard triplet loss applied on the visible image set $\boldsymbol{V}$ and infrared image set $\boldsymbol{I}$ can be formulated as follows:

$$\mathcal{L}_{triplet}^{\boldsymbol{VI}} = \sum_{\boldsymbol{f}^a, \boldsymbol{f}^p, \boldsymbol{f}^n \in \boldsymbol{F^{VI}}} [d(\boldsymbol{f}^a, \boldsymbol{f}^p) - d(\boldsymbol{f}^a, \boldsymbol{f}^n) + m]_+ \tag{2}$$

where $\boldsymbol{F^{VI}} = [\boldsymbol{F^V}||\boldsymbol{F^I}] = [\boldsymbol{f}_1^{\boldsymbol{V}}, \cdots, \boldsymbol{f}_N^{\boldsymbol{V}}, \boldsymbol{f}_1^{\boldsymbol{I}}, \cdots, \boldsymbol{f}_N^{\boldsymbol{I}}]$, $||$ denotes the feature vector sets union operation. $d(\cdot,\cdot)$ calculates the euclidean distance. m is the margin parameter. Following (Hermans, Beyer, and Leibe 2017), for each anchor $\boldsymbol{f}^a$ in $\boldsymbol{F^{VI}}$, we select the hardest positive sample $\boldsymbol{f}^p$ (furthest distance with anchor) and the hardest negative sample $\boldsymbol{f}^n$ (nearest distance with anchor) within $\boldsymbol{F^{VI}}$. In addition, we also impose the Kullback-Leibler (KL) divergence constraint to encourage the similarity of the predicted probability distribution for the same identity from two modalities (Hao et al. 2019), and apply a bi-directional KL loss to achieve such constraint.

$$\mathcal{L}_{KL}^{\boldsymbol{VI}} = \mathbb{E}_i \sum_{c=0}^{C-1} [p(c|\boldsymbol{f}_i^{\boldsymbol{V}}) \log \frac{p(c|\boldsymbol{f}_i^{\boldsymbol{V}})}{p(c|\boldsymbol{f}_i^{\boldsymbol{I}})} + p(c|\boldsymbol{f}_i^{\boldsymbol{I}}) \log \frac{p(c|\boldsymbol{f}_i^{\boldsymbol{I}})}{p(c|\boldsymbol{f}_i^{\boldsymbol{V}})}] \tag{3}$$

Color-irrelevant Consistency Learning

Since the features of visible images tend to overfit the clothing color information (Zheng et al. 2019b), which becomes an obstacle to effective modality alignment in VI-ReID task. In light of this, we present the color-irrelevant consistency learning (CICL) to learn the color-irrelevant features so that the heterogeneous feature distributions can be better adapted and modality gap could be decreased.

To learn color-irrelevant features, a naive idea is to convert RGB images to grayscale images to eliminate the color cues. However, image graying may lose discriminative information. Meanwhile, simple image graying is a relatively weak constraint to learn color-irrelevant features. Instead of eliminating the color cues, the CICL attempts to introduce random clothing color transformation to visible images and further imposes much stronger color-irrelevant consistency constraints to learn color-irrelevant features.

Clothing Color Transformation. To generate RGB images which share the same identity but with different colored clothes, we introduce our clothing color transformation method. As shown in Figure 2, for each input visible image in $\boldsymbol{V}$, we first adopt the human semantic parser SCHP model (Li et al. 2019) to obtain the corresponding human parsing mask. Then, we selectively divide the mask into two major categories: clothing region (upper cloth, pants, shoes, etc) and non-clothing region (face, arm, leg, and background). Next we utilize the color jitter operation to randomly adjust the hue, saturation, contrast and brightness of clothing region and maintain non-clothing region unchanged. Thus in each mini-batch or different training iterations, an identity will wear diverse color clothes. Finally, the corresponding clothing color transformed image set $\boldsymbol{T} = \{T_i\}_{i=1}^{N}$ is derived and the identity label set of $\boldsymbol{T}$ is naturally the same as $\boldsymbol{V}$. Similar to $\boldsymbol{V}$, $\boldsymbol{T}$ is also fed into the RGB branch to extract the corresponding feature vector set $\boldsymbol{F^T} = [\boldsymbol{f}_1^{\boldsymbol{T}}, \boldsymbol{f}_2^{\boldsymbol{T}}, \cdots, \boldsymbol{f}_i^{\boldsymbol{T}}, \cdots, \boldsymbol{f}_N^{\boldsymbol{T}}]$.

Intra-modality Color-irrelevant Consistency Learning. To enhance the discriminative color-irrelevant feature learning, we first impose the intra-modality color-irrelevant consistency constraint on visible image set $\boldsymbol{V}$ and the corresponding clothing color transformed image set $\boldsymbol{T}$.

The first is the intra-modality color-irrelevant identity prediction consistency, which requires that even if a person changes the clothing color, the model should make consistent identity prediction. It enforces model not to focus too much on specific clothes colors and pay more attention to learn the body shape, clothing style and other color-irrelevant features to distinguish identities. We apply identity loss on $\boldsymbol{T}$ to achieve above constraint, where $y_i^{\boldsymbol{T}} = y_i^{\boldsymbol{V}}$.

$$\mathcal{L}_{identity}^{\boldsymbol{T}} = \mathbb{E}_{i,\boldsymbol{T}}[-\log(p(y_i^{\boldsymbol{T}}|\boldsymbol{f}_i^{\boldsymbol{T}}))] \tag{4}$$

The second is the intra-modality color-irrelevant feature embedding consistency. It encourages that for the same identity within RGB modality, no matter what color clothing he/she wears, the feature distances between them should be as small as possible. It also enforces that for different identities, even if they are wearing clothes of similar colors, the feature distances between them should be as large as possible. We impose the hard triplet loss on $\boldsymbol{F^{VT}}$ to achieve above constraint, where the $\boldsymbol{F^{VT}} = [\boldsymbol{F^V}||\boldsymbol{F^T}]$.

$$\mathcal{L}_{triplet}^{\boldsymbol{VT}} = \sum_{\boldsymbol{f}^a,\boldsymbol{f}^p,\boldsymbol{f}^n \in \boldsymbol{F^{VT}}} [d(\boldsymbol{f}^a,\boldsymbol{f}^p) - d(\boldsymbol{f}^a,\boldsymbol{f}^n) + m]_+ \tag{5}$$

The third is intra-modality color-irrelevant probability distribution consistency. It imposes the similarity constraint to the predicted probability distributions of visible image pair composed of the RGB images before and after transformation, which further strengthens the color-irrelevant constraint and facilitates the color-irrelevant feature learning. We apply the bi-directional KL divergence loss between $\boldsymbol{V}$ and $\boldsymbol{T}$ to impose above constraint.

$$\mathcal{L}_{KL}^{\boldsymbol{VT}} = \mathbb{E}_i \sum_{c=0}^{C-1}[p(c|\boldsymbol{f}_i^{\boldsymbol{V}})\log\frac{p(c|\boldsymbol{f}_i^{\boldsymbol{V}})}{p(c|\boldsymbol{f}_i^{\boldsymbol{T}})} + p(c|\boldsymbol{f}_i^{\boldsymbol{T}})\log\frac{p(c|\boldsymbol{f}_i^{\boldsymbol{T}})}{p(c|\boldsymbol{f}_i^{\boldsymbol{V}})}] \tag{6}$$

Inter-modality Color-irrelevant Consistency Learning. Although the intra-modality CICL effectively facilitates the color-irrelevant feature learning, it is not sufficient to only impose the consistency constraint within RGB modality. To this end, we also introduce complementary inter-modality color-irrelevant consistency learning, which encourages to learn complementary cross modality color-irrelevant features and enhance discrimination.

We first impose the inter-modality color-irrelevant feature embedding consistency constraint. It encourages that for the visible images which share the same identity but with different clothing colors, they should close to the infrared images that with the same identity. We apply the hard triplet loss on $\boldsymbol{F^{TI}}$ to achieve above constraint, where $\boldsymbol{F^{TI}} = [\boldsymbol{F^T}||\boldsymbol{F^I}]$.

$$\mathcal{L}_{triplet}^{\boldsymbol{TI}} = \sum_{\boldsymbol{f}^a,\boldsymbol{f}^p,\boldsymbol{f}^n \in \boldsymbol{F^{TI}}} [d(\boldsymbol{f}^a,\boldsymbol{f}^p) - d(\boldsymbol{f}^a,\boldsymbol{f}^n) + m]_+ \tag{7}$$

We further apply the inter-modality color-irrelevant probability distribution consistency constraint. It demands that for the visible images with the same identity, no matter what color clothes they wear, the predicted probability distributions of them should be as similar as possible to that of the same identity in IR modality. We apply the bi-directional KL divergence loss between $\boldsymbol{T}$ and $\boldsymbol{I}$ to impose above constraint as follows:

$$\mathcal{L}_{KL}^{\boldsymbol{TI}} = \mathbb{E}_i \sum_{c=0}^{C-1}[p(c|\boldsymbol{f}_i^{\boldsymbol{T}})\log\frac{p(c|\boldsymbol{f}_i^{\boldsymbol{T}})}{p(c|\boldsymbol{f}_i^{\boldsymbol{I}})} + p(c|\boldsymbol{f}_i^{\boldsymbol{I}})\log\frac{p(c|\boldsymbol{f}_i^{\boldsymbol{I}})}{p(c|\boldsymbol{f}_i^{\boldsymbol{T}})}] \tag{8}$$

Identity-aware Modality Adaptation

Learning color-irrelevant features could decrease the modality gap and promote modality alignment. On top of this, we propose the identity-aware modality adaptation (IAMA) to further align the identity-level feature distributions of RGB and IR modality into a same space, which aims to tackle the problem of feature misalignment of some identities resulting from dataset-level modality alignment and retain the discriminative ability of the learned features. By embedding cross modality features into a fine-grained modality-shared space, the IAMA in turn promotes the two-stream feature extractor to learn discriminative color-irrelevant features. In this paper, we adpot the widely-used domain discrepancy metric Maximum Mean Discrepancy (MMD) (Gretton et al. 2012a) to adapt the heterogeneous feature distributions, below we first briefly revisit the MMD loss.

Suppose there are two sample sets $\mathcal{A} = \{\boldsymbol{a}_i\}_{i=1}^{n_A}$ and $\mathcal{B} = \{\boldsymbol{b}_j\}_{j=1}^{n_B}$, where $\boldsymbol{a}_i$ and $\boldsymbol{b}_j$ are feature vectors sampled from two different distributions. Specifically, the MMD loss calculates the kernel mean embedding distance of two distributions in Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ associated with the kernel function $k(\cdot,\cdot)$. Formally, the squared MMD loss can be formulated as follows:

$$\mathcal{L}_{\mathrm{MMD}^2}(\mathcal{A},\mathcal{B}) \triangleq \left\|\mathbf{E}_a\left[\phi\left(\boldsymbol{a}\right)\right] - \mathbf{E}_b\left[\phi\left(\boldsymbol{b}\right)\right]\right\|^2_{\mathcal{H}_k} \tag{9}$$

where $\phi(\cdot)$ is an implicit feature mapping function. By applying the kernel trick $k(\boldsymbol{a}_i, \boldsymbol{b}_j) = \phi(\boldsymbol{a}_i)^T\phi(\boldsymbol{b}_j)$, the Eq. (9) can be simplified and calculated. In our paper, we adopt the Multi-Kernel MMD (Gretton et al. 2012b) with gaussian kernel.

In VI-ReID task, some works (Dai et al. 2018; Lu et al. 2020) leverage adversarial learning to align the feature distributions of RGB and IR modality in dataset-level while ignoring the valuable identity information, which may cause the features of some identities are misaligned. In contrast, the proposed IAMA fully exploits the built-in identity labels and thus align the identity-level cross modality distributions. To implement identity-level adaptation, we introduce batch sampling strategy. Specially, in a training batch, P person identities are randomly selected, and then we randomly select K visible images and K infrared images for each selected identity. We further group the images within a batch by identity label and calculate the identity-level MMD loss rather than dataset-level MMD loss. Hence, the IAMA imposed between $\boldsymbol{V}$ and $\boldsymbol{I}$ could be formulated as follows:

$$\mathcal{L}_{IAMA}(\boldsymbol{V},\boldsymbol{I}) = \frac{1}{P}\sum_{p=1}^{P}\left\|\mathbf{E}\left[\phi\left(\boldsymbol{F}_p^{\boldsymbol{V}}\right)\right] - \mathbf{E}\left[\phi\left(\boldsymbol{F}_p^{\boldsymbol{I}}\right)\right]\right\|^2_{\mathcal{H}_k} \tag{10}$$

where $\boldsymbol{F}_p^{\boldsymbol{V}}$ and $\boldsymbol{F}_p^{\boldsymbol{I}}$ indicate the feature vector sets belonging to the p^{th} identity in $\boldsymbol{F}^{\boldsymbol{V}}$ and $\boldsymbol{F}^{\boldsymbol{I}}$, respectively. $\mathbf{E}$ in Eq. (10) is to calculate the kernel mean embedding of features in $\boldsymbol{F}_p^{\boldsymbol{V}}$ and $\boldsymbol{F}_p^{\boldsymbol{I}}$, i.e, $\mathbf{E}\left[\phi\left(\boldsymbol{F}_p^{\boldsymbol{V}}\right)\right] = \frac{1}{K}\sum_{k=1}^{K}\phi(\boldsymbol{f}_k), \boldsymbol{f}_k \in \boldsymbol{F}_p^{\boldsymbol{V}}$, $\phi(\cdot)$ is an implicit function to map features to RKHS space $\mathcal{H}_k$. Meanwhile, since both $\boldsymbol{T}$ and $\boldsymbol{V}$ belong to the RGB modality, we also impose the identity-aware modality adaptation between $\boldsymbol{T}$ and $\boldsymbol{I}$. The entire IAMA loss is formulated as follows:

$$\mathcal{L}^{\boldsymbol{VTI}}_{IAMA} = \mathcal{L}_{IAMA}(\boldsymbol{V},\boldsymbol{I}) + \mathcal{L}_{IAMA}(\boldsymbol{T},\boldsymbol{I}) \tag{11}$$

Optimization

We mix the loss functions into the following form according to the type of loss functions.

$$\mathcal{L}^{\boldsymbol{VTI}}_{identity} = \mathbb{E}_{i,\boldsymbol{M}}[-\log(p(y_i^{\boldsymbol{M}}|\boldsymbol{f}_i^{\boldsymbol{M}})], \boldsymbol{M} \in \{\boldsymbol{V},\boldsymbol{T},\boldsymbol{I}\} \tag{12}$$

$$\mathcal{L}^{\boldsymbol{VTI}}_{triplet} = \mathcal{L}^{\boldsymbol{VI}}_{triplet} + \mathcal{L}^{\boldsymbol{VT}}_{triplet} + \mathcal{L}^{\boldsymbol{TI}}_{triplet} \tag{13}$$

$$\mathcal{L}^{\boldsymbol{VTI}}_{KL} = \mathcal{L}^{\boldsymbol{VI}}_{KL} + \mathcal{L}^{\boldsymbol{VT}}_{KL} + \mathcal{L}^{\boldsymbol{TI}}_{KL} \tag{14}$$

We assign equal importance to identity loss and triplet loss and use λ_1 and λ_2 to control the weights of KL loss and IAMA loss. When calculating the triplet loss and IAMA loss in this paper, we use l_2-normalized feature vectors. In summary, the overall loss is as follows:

$$\mathcal{L} = \mathcal{L}^{\boldsymbol{VTI}}_{identity} + \mathcal{L}^{\boldsymbol{VTI}}_{triplet} + \lambda_1\mathcal{L}^{\boldsymbol{VTI}}_{KL} + \lambda_2\mathcal{L}^{\boldsymbol{VTI}}_{IAMA} \tag{15}$$

We optimize the entire network in an end-to-end manner by minimizing the overall loss function.

Experiments

Experimental Settings

Datasets SYSU-MM01 (Wu et al. 2017) is a large-scale VI-ReID dataset captured by 6 cameras, including 4 visible and 2 near-infrared cameras. The training set contains 22,258 RGB images and 11,909 infrared images of 395 persons. The testing set includes 96 persons with 3,803 infrared images as query and 301/3010 (single shot/multi-shot) randomly selected RGB images as gallery. There are two testing mode: all-search and indoor-search mode. For all-search mode, the gallery consists of the RGB images captured in both indoor and outdoor environments, while the gallery set in indoor-search mode is only composed of the RGB images captured by indoor cameras. The experiments would run 10 times to get average performance following (Wu et al. 2017).

RegDB (Nguyen et al. 2017) is constructed by dual camera and includes 412 persons. For each person, 10 visible images are captured by a visible camera, and 10 infrared images are obtained by a far-infrared camera. We follow the evaluation protocol in (Ye et al. 2018a,b) , where the dataset is randomly split into two halves, one for training and the other for testing. The procedure is repeated for 10 trials to obtain the average performance. For both datasets, the Rank-k accuracy and mean Average Precision (mAP) are adopted to evaluate the performance following (Wu et al. 2017).

Implementation Details. Our approach is implemented with PyTorch framework on one NVIDIA Titan Xp GPU. We adopt the ResNet50 (He et al. 2016)model pretrained on ImageNet as the backbone network. Following (Luo et al. 2019), we remove the last stride of ResNet50. The visible and infrared images are resized to 3 × 288×144. For infrared images, the three channels are the same. During training, we adopt the random horizontal flip and random erasing (Zhong et al. 2020) for data augmentation follows (Luo et al. 2019). For each mini-batch, 4 identities are randomly selected and we randomly sample 8 visible images and 8 infrared images for each selected identity. We use the SGD algorithm as the optimizer, then we train the model for 90 epochs with base learning rate initialized at 0.03 and decaying 10 times at 45, 70 epoch. The learning rate for all pretrained layers is set to 0.1 times of the base learning rate. The λ_1 is set to 0.5 for both datasets, the λ_2 is empirically set to 0.3 and 0.7 for SYSU-MM01 and RegDB respectively, since there is a larger modality gap in RegDB (far-infrared) compared with SYSU-MM01 (near-infrared). When performing clothing color transformation of visible images, the input 4 configuration arguments of color jitter are all set to 0.3 for SYSU-MM01, set to 0.5 for RegDB. During testing, for images in query and gallery, we extract the l_2-normalized 512-dimensional feature vectors from corresponding modality branch and use euclidean distance to rank.

Comparison with State-of-the-art Methods

We compare our approach with the state-of-the-art methods in VI-ReID. The comparison results on SYSU-MM01 dataset are shown in Table 1. In the most challenging single-shot all search mode, our approach achieves **59.3%** mAP and **57.2%** Rank-1 on SYSU-MM01 dataset, which signif-

Methods		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
Zero-Padding (Wu et al. 2017)		14.8	54.1	71.3	15.9	19.1	61.4	78.4	10.9	20.6	68.4	85.8	26.9	24.4	75.9	91.3	18.6
TONE (Ye et al. 2018a)		12.5	50.7	68.6	14.4	-	-	-	-	-	-	-	-	-	-	-	-
BCTR (Ye et al. 2018b)		16.1	54.9	71.5	19.1	-	-	-	-	-	-	-	-	-	-	-	-
BDTR (Ye et al. 2018b)		17.0	55.4	72.0	19.7	-	-	-	-	-	-	-	-	-	-	-	-
D-HSME (Hao et al. 2019)		20.7	62.8	78.0	23.2	-	-	-	-	-	-	-	-	-	-	-	-
cmGAN (Dai et al. 2018)		27.0	67.5	80.6	27.8	31.5	72.7	85.0	22.3	31.6	77.2	89.2	42.2	37.0	80.9	92.1	32.8
D^{2}RL (Wang et al. 2019b)		28.9	70.6	82.4	29.2	-	-	-	-	-	-	-	-	-	-	-	-
MAC (Ye, Lan, and Leng 2019)		33.2	79.0	90.0	36.2	-	-	-	-	33.3	82.4	93.6	44.9	-	-	-	-
Hi-CMD (Choi et al. 2020)		34.9	-	77.5	35.9	-	-	-	-	-	-	-	-	-	-	-	-
JSIA-ReID (Wang et al. 2020)		38.1	80.7	89.9	36.9	45.1	85.7	93.8	29.5	43.8	86.2	94.2	52.9	52.7	91.1	96.4	42.7
AlignGAN (Wang et al. 2019a)		42.4	85.0	93.7	40.7	51.5	89.4	95.7	33.9	45.9	87.6	94.4	54.3	57.1	92.7	97.4	45.3
FMSP (Wu et al. 2020)		43.6	74.6	86.2	44.9	-	-	-	-	48.7	79.0	89.5	57.5	-	-	-	-
DFE (Hao et al. 2019)		48.7	88.8	95.2	48.5	54.6	91.6	96.8	42.1	52.2	89.8	95.8	59.6	59.6	94.4	98.0	50.6
XIV-ReID (Li et al. 2020)		49.9	89.7	95.9	50.7	-	-	-	-	-	-	-	-	-	-	-	-
DDAG (Ye et al. 2020)		54.7	90.3	95.8	53.0	-	-	-	-	61.0	94.6	98.4	67.9	-	-	-	-
HAT (Ye, Shen, and Shao 2020)		55.2	92.4	97.3	53.8	-	-	-	-	62.1	95.7	99.2	69.3	-	-	-	-
cm-SSFT (Lu et al. 2020)	Single query	47.7	-	-	54.1	57.4	-	-	59.1	-	-	-	-	-	-	-	-
	All queries	61.6	89.2	93.9	63.2	63.4	91.2	95.7	62.0	70.5	94.9	97.7	72.6	73.0	96.3	99.1	72.4
Ours		**57.2**	94.3	98.4	**59.3**	**60.7**	95.2	98.6	**52.6**	**66.6**	98.8	99.7	**74.7**	**73.8**	99.4	99.9	**68.3**

Table 1: Comparison with the state-of-the-arts on SYSU-MM01 dataset. The R1, R10, R20 denote Rank-1, 10 and 20 accuracies (%), respectively. The mAP denotes mean average precision score (%).

Methods		V → I		I→ V	
		R1	mAP	R1	mAP
Zero-Padding (Wu et al. 2017)		17.8	18.9	16.7	17.9
TONE (Ye et al. 2018a)		16.9	14.9	13.9	17.0
BCTR (Ye et al. 2018b)		32.7	31.0	-	-
BDTR (Ye et al. 2018b)		33.5	31.8	32.7	31.1
MAC (Ye, Lan, and Leng 2019)		36.4	37.0	36.2	36.6
D^{2}RL (Wang et al. 2019b)		43.4	44.1	-	-
JSIA-ReID (Wang et al. 2020)		48.5	49.3	48.1	48.9
D-HSME (Hao et al. 2019)		50.9	47.0	50.2	46.2
AlignGAN (Wang et al. 2019a)		57.9	53.6	56.3	53.4
XIV-ReID (Li et al. 2020)		-	-	62.2	60.1
FMSP (Wu et al. 2020)		65.0	64.5	-	-
DDAG (Ye et al. 2020)		69.3	63.4	68.0	61.8
DFE (Hao et al. 2019)		70.1	69.1	67.9	66.7
Hi-CMD (Choi et al. 2020)		70.9	66.0	-	-
HAT(Ye, Shen, and Shao 2020)		71.8	67.5	70.0	66.3
cm-SSFT (Lu et al. 2020)	Single query	65.4	65.5	63.8	64.2
	All queries	72.3	72.9	71.0	71.7
Ours		**78.8**	69.4	**77.9**	69.4

Table 2: Comparison with the state-of-the-arts on the RegDB dataset. V → I means visible-search-infrared mode, while I → V means the opposite mode.

Methods	SYSU-MM01		RegDB	
	Rank-1	mAP	Rank-1	mAP
baseline	45.3	46.7	52.3	51.8
baseline + CICL	53.0	54.8	62.1	60.1
baseline + IAMA	52.6	53.5	70.0	60.8
baseline + CICL + IAMA	57.2	59.3	78.8	69.4

Table 3: The effectiveness of our proposed color-irrelevant consistency learning (CICL) and identity-aware modality adaptation (IAMA).

Methods	SYSU-MM01	
	Rank-1	mAP
baseline	45.3	46.7
baseline + intra-modality CICL	51.4	51.9
baseline + inter-modality CICL	49.2	49.7
baseline + CICL (intra+inter)	53.0	54.8

Table 4: The comparison of intra and inter modality CICL.

icantly outperforms the state-of-the-art methods, including DDAG (Ye et al. 2020), HAT (Ye, Shen, and Shao 2020) and cm-SSFT (Lu et al. 2020). Note that we compare the single query version of cm-SSFT for fairness, since all queries version of cm-SSFT utilizes auxiliary set for testing, which is prohibited in formal test protocol. In other evaluation modes, i.e., all-search / indoor-search and single-shot / multi-shot mode, our approach also consistently exceeds the performance of state-of-the-arts, which demonstrates the robustness and effectiveness of our method. The comparison results on RegDB dataset are shown in Table 2. We achieves the **78.8%** Rank-1 and **69.4%** mAP under V → I mode, which outperforms the state-of-the-arts including Hi-CMD, HAT and cm-SSFT by a large margin. Our methods also achieves state-of-the-art performance in I → V mode.

Compared with the early GAN-based methods (Wang et al. 2019b; Dai et al. 2018; Wang et al. 2019a, 2020), our methods does not introduce any additional learnable parameters, nor does it add extra computational cost during testing, which is more efficient and effective. Compared with auxiliary modality augmented methods (Li et al. 2020; Ye, Shen, and Shao 2020), our methods could learn more discriminative and modality-shared features, which achieves much better performance on two VI-ReID datasets. Compared with recent proposed Hi-CMD and cm-SSFT, our method does not leverage complicated disentanglement strategy, and performs identity-aware modality adaptation, which would be more suitable when deployed in pratical scenario.

Ablation Study

The Effectiveness of the Proposed CICL and IAMA. As shown in Table 3, compared with the two stream baseline, the CICL brings 8.1% and 8.3% mAP improvements on SYSU-MM01 and RegDB respectively, which fully demon-

consistency	identity prediction	feature embedding	probability distribution	SYSU-MM01	
				Rank-1	mAP
baseline	–	–	–	45.3	46.7
1	✓	–	–	48.9	49.6
2	✓	✓	–	51.1	52.3
3	✓	✓	✓	53.0	54.8

Table 5: The effectiveness of various color-irrelevant consistency constraints.

Methods	RegDB	
	Rank-1	mAP
baseline + Grayscaling	55.8	53.3
baseline + ColorAugmentation	58.1	56.2
baseline + CICL	62.1	60.1

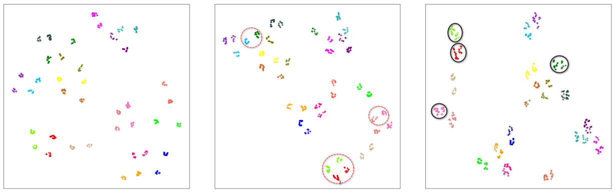
Table 6: The comparison of the color-irrelevant consistency learning with auxiliary grayscale image and naive color augmentation on the baseline.

strates that learning the color-irrelevant features could effectively decrease the modality gap. We can also observe that the IAMA brings 17.7% Rank-1 gain to the baseline on RegDB dataset, which highlights the effectiveness and necessity of identity-level modality adaptation. More importantly, when incorporating the CICL and IAMA together, the performance is further improved, which shows that color-irrelevant feature learning and identity-aware modality adaptation are complementary and can promote each other, removing any one will result in suboptimal performance.

The Analysis of Color-irrelevant Consistency Learning. Since the CICL integrates both intra and inter modality CICL, we also conduct ablation study to verify the effectiveness of them. As shown in the Table 4, both intra modality and inter modality CICL could boost the performance. Among them, intra-modality CICL is more effective and inter-modality CICL plays a complementary role. Furthermore, we also analyze the effectiveness of various kinds of color-irrelevant consistency constraints. As shown in Tab. 5, on the baseline, we progressively impose the identity prediction consistency constraint, feature embedding consistency (both intra+inter modality) constraint and probability distribution consistency (both intra+inter modality) constraint, the corresponding performance is also gradually improved, which highlights the complementary and effectiveness of above consistency constraints.

We further compare the proposed CICL with auxiliary image graying (replace $\boldsymbol{T}$ to grayscale image) and naive color augmentation (baseline with color augmentation). As shown in Tab. 6, the CICL significantly outperforms auxiliary image graying and direct color augmentation on the baseline. Since auxiliary grayscale image may lose some discriminative information and imposes relatively weak color-irrelevant constraint, while the baseline with naive color augmentation does not impose such a strong intra-modality color-irrelevant constraints and it is not just transforming the clothing region color, which may damage the non-clothing region (such as face) and harm the performance.

The Analysis of the Identity-aware Modality Adaptation.



(a) baseline (bs) (b) bs+dataset-level (c) bs+identity-level

Figure 3: The t-SNE visualization of features with identity-level (IAMA) and dataset-level modality adaptation based on baseline. 20 identities are randomly selected from RegDB testset. Samples with same color indicate the same identity. The markers “dot” and “cross” denote RGB and IR modality.

Methods	RegDB	
	Rank-1	mAP
baseline	52.3	51.8
baseline + dataset-level	64.3	54.1
baseline + identity-level	70.0	60.8

Table 7: The comparison of identity-level alignment (IAMA) with dataset-level modality alignment.

We further verify the effectiveness of identity-aware modality adaptation (IAMA) both qualitatively and quantitatively. For comparison, we conduct both identity-level and dataset-level alignment on our baseline. As shown in Table 7, compare with the baseline, both identity-level and dataset-level modality alignment can bring large performance improvements to. More importantly, we can observe that identity-aware alignment significantly surpasses the performance of dataset-level alignment. Quantitative comparison is also conducted, we utilize t-SNE to visualize the features in 2-D plane. As shown in Fig. 3(a), the initial distance between cross modality features of the same identity is relatively large. After the modality alignment, the modality gap between the RGB and IR modality are effectively decreased, as shown in Fig. 3(b) and (c). However, for dataset-level modality alignment shown in Fig. 3(b), we can observe that the features of some identities are misaligned (marked with red circles), which may harm the discrimination of features. In contrast, as shown in Fig. 3(c), by applying the proposed identity-level modality adaptation, the features of misaligned identities in Fig. 3(b) are now correctly aligned (marked with black circles), which further proves the necessity of aligning cross modality features in identity-level.

Conclusion

In this paper, we propose a novel approach for visible-infrared cross modality person re-identification. To effectively decrease the modality gap, our approach learns the color-irrelevant features through the color-irrelevant consistency learning and aligns the identity-level feature distributions by the identity-aware modality adaptation. Extensive experiments demonstrate that our approach significantly surpasses the state-of-the-art methods and ablation studies further validate the effectiveness of each component.

References

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *NIPS*, 5049–5059.

Chen, W.; Chen, X.; Zhang, J.; and Huang, K. 2017. Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-identification. In *CVPR*, 1320–1329.

Choi, S.; Lee, S.; Kim, Y.; Kim, T.; and Kim, C. 2020. Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification. In *CVPR*, 10257–10266.

Dai, P.; Ji, R.; Wang, H.; Wu, Q.; and Huang, Y. 2018. Cross-Modality Person Re-Identification with Generative Adversarial Training. In *IJCAI*, 677–683.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; SchÃ¶lkopf, B.; and Smola, A. 2012a. A kernel two-sample test. *Journal of Machine Learning Research* 13(1): 723–773.

Gretton, A.; Sejdinovic, D.; Strathmann, H.; Balakrishnan, S.; Pontil, M.; Fukumizu, K.; and Sriperumbudur, B. K. 2012b. Optimal kernel choice for large-scale two-sample tests. In *NIPS*, volume 2, 1205–1213.

Hao, Y.; Wang, N.; Gao, X.; Li, J.; and Wang, X. 2019. Dual-alignment Feature Embedding for Cross-modality Person Re-identification. In *ACM MM*.

Hao, Y.; Wang, N.; Jie, L.; and Gao, X. 2019. HSME: Hypersphere Manifold Embedding for Visible Thermal Person Re-identification. In *AAAI*, 8385–8392.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737* .

Kniaz, V. V.; Knyaz, V. A.; Hladuvka, J.; Kropatsch, W. G.; and Mizginov, V. 2018. ThermalGAN: Multimodal Color-to-Thermal Image Translation for Person Re-identification in Multispectral Dataset. In *ECCV*, 606–624.

Laine, S.; and Aila, T. 2017. Temporal Ensembling for Semi-Supervised Learning. In *ICLR*.

Li, D.; Wei, X.; Hong, X.; and Gong, Y. 2020. Infrared-Visible Cross-Modal Person Re-Identification with an X Modality. In *AAAI*, 4610–4617.

Li, P.; Xu, Y.; Wei, Y.; and Yang, Y. 2019. Self-Correction for Human Parsing. *arXiv preprint arXiv:1910.09777* .

Liu, F.; and Zhang, L. 2019. View Confusion Feature Learning for Person Re-Identification. In *ICCV*, 6638–6647.

Lu, Y.; Wu, Y.; Liu, B.; Zhang, T.; Li, B.; Chu, Q.; and Yu, N. 2020. Cross-modality Person re-identification with Shared-Specific Feature Transfer. In *CVPR*, 13379–13389.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Nguyen, D. T.; Hong, H. G.; Kim, K. W.; and Park, K. R. 2017. Person Recognition System Based on a Combination of Body Images from Visible Light and Thermal Cameras. *Sensors* 17(3): 605.

Qian, X.; Fu, Y.; Xiang, T.; Wang, W.; Qiu, J.; Wu, Y.; Jiang, Y.-G.; and Xue, X. 2018. Pose-Normalized Image Generation for Person Re-identification. In *ECCV*, 661–678.

Sajjadi, M.; Javanmardi, M.; and Tasdizen, T. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NIPS*, 1171–1179.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In *ECCV*, 501–518.

Varior, R. R.; Haloi, M.; and Wang, G. 2016. Gated Siamese Convolutional Neural Network Architecture for Human Re-identification. In *ECCV*, 791–808.

Wang, G.; Zhang, T.; Cheng, J.; Liu, S.; Yang, Y.; and Hou, Z. 2019a. RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment. In *ICCV*, 3622–3631.

Wang, G.-A.; Zhang, T.; Yang, Y.; Cheng, J.; Chang, J.; Liang, X.; and Hou, Z. 2020. Cross-Modality Paired-Images Generation for RGB-Infrared Person Re-Identification. In *AAAI*.

Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.-Y.; and Satoh, S. 2019b. Learning to Reduce Dual-Level Discrepancy for Infrared-Visible Person Re-Identification. In *CVPR*, 618–626.

Wei, L.; Zhang, S.; Yao, H.; Gao, W.; and Tian, Q. 2017. GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval. In *2017 ACM on Multimedia Conference*, 420–428.

Wu, A.; Zheng, W.-S.; Gong, S.; and Lai, J. 2020. RGB-IR Person Re-identification by Cross-Modality Similarity Preservation. *IJCV* 1–21.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. RGB-Infrared Cross-Modality Person Re-identification. In *ICCV*, 5390–5399.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.-T.; and Le, Q. V. 2019. Unsupervised Data Augmentation for Consistency Training. *arXiv preprint arXiv:1904.12848* .

Yang, Q.; Wu, A.; and Zheng, W.-S. 2020. Person Re-identification by Contour Sketch under Moderate Clothing Change. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1.

Ye, M.; Lan, X.; and Leng, Q. 2019. Modality-aware Collaborative Learning for Visible Thermal Person Re-Identification. In *ACM MM*.

Ye, M.; Lan, X.; Li, J.; and c Yuen, P. 2018a. Hierarchical Discriminative Learning for Visible Thermal Person Re-Identification. In *AAAI*, 7501–7508.

Ye, M.; Shen, J.; Crandall, D. J.; Shao, L.; and Luo, J. 2020. Dynamic Dual-Attentive Aggregation Learning for Visible-Infrared Person Re-Identification. In *ECCV*.

Ye, M.; Shen, J.; and Shao, L. 2020. Visible-Infrared Person Re-Identification via Homogeneous Augmented Tri-Modal Learning. *IEEE Transactions on Information Forensics and Security* 1–1.

Ye, M.; Wang, Z.; Lan, X.; and Yuen, P. C. 2018b. Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking. In *IJCAI*, 1092–1099.

Zheng, L.; Huang, Y.; Lu, H.; and Yang, Y. 2019a. Pose-Invariant Embedding for Deep Person Re-Identification. *IEEE Transactions on Image Processing* 28(9): 4500–4509.

Zheng, L.; Yang, Y.; and Hauptmann, A. G. 2016. Person Re-identification: Past, Present and Future. *arXiv preprint arXiv:1610.02984* .

Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; and Kautz, J. 2019b. Joint Discriminative and Generative Learning for Person Re-Identification. In *CVPR*, 2138–2147.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random Erasing Data Augmentation. In *AAAI*.