# Modality-Adaptive Mixup and Invariant Decomposition for RGB-Infrared Person Re-Identification

**Zhipeng Huang**[1*], **Jiawei Liu**[1*], **Liang Li**[2], **Kecheng Zheng**[1], **Zheng-Jun Zha**[1†]

[1] University of Science and Technology of China
[2] Institute of Computing Technology, Chinese Academy of Sciences
{hzp1104,zkcys001}@mail.ustc.edu.cn, {jwliu6,zhazj}@ustc.edu.cn, liang.li@ict.ac.cn

## Abstract

RGB-infrared person re-identification is an emerging cross-modality re-identification task, which is very challenging due to significant modality discrepancy between RGB and infrared images. In this work, we propose a novel modality-adaptive mixup and invariant decomposition (MID) approach for RGB-infrared person re-identification towards learning modality-invariant and discriminative representations. MID designs a modality-adaptive mixup scheme to generate suitable mixed modality images between RGB and infrared images for mitigating the inherent modality discrepancy at the pixel-level. It formulates modality mixup procedure as Markov decision process, where an actor-critic agent learns dynamical and local linear interpolation policy between different regions of cross-modality images under a deep reinforcement learning framework. Such policy guarantees modality-invariance in a more continuous latent space and avoids manifold intrusion by the corrupted mixed modality samples. Moreover, to further counter modality discrepancy and enforce invariant visual semantics at the feature-level, MID employs modality-adaptive convolution decomposition to disassemble a regular convolution layer into modality-specific basis layers and a modality-shared coefficient layer. Extensive experimental results on two challenging benchmarks demonstrate superior performance of MID over state-of-the-art methods.

## Introduction

Person re-identification (Re-ID) has attracted increasing attention recently, due to its widespread applications in automated tracking (Wang 2013; Kim et al. 2017) and activity analysis (Aggarwal and Ryoo 2011; Caba Heilbron et al. 2015), *etc*. It aims at identifying a target pedestrian from a gallery set captured from multi-disjoint camera views. Person Re-ID is quite challenging due to large intra-class and small inter-class variations caused by background clutter, occlusion, dramatic variations in illumination, body pose, *etc*. Most existing person Re-ID methods mainly focus on RGB images of pedestrians from visible cameras and formulate the task as a single-modality (RGB-RGB) matching problem. They have achieved remarkable progresses in recent years for addressing appearance discrepancy (large intra-class and small inter-class variations). However, visible (RGB) cameras can not provide useful appearance information under poor illumination environments (*e.g.*, at night), which limits the applicability of person Re-ID in a real scenario.

To handle this issue, recent surveillance systems begin to be equipped with infrared (IR) cameras to facilitate nighttime monitoring, which raises a new cross-modality matching task termed RGB-infrared person Re-ID (Wu et al. 2017). RGB-infrared person Re-ID aims to find the corresponding IR (or RGB) images of the same person captured by other spectrum cameras, given a RGB (or IR) image of a target person. Compared with the conventional single-modality person Re-ID, it encounters prominent modality discrepancy derived from the different imaging processes between different spectrum cameras (RGB and infrared images are intrinsically heterogeneous, which have different wavelength ranges), apart from appearance discrepancy. The key solution for RGB-infrared person Re-ID is to bridge large modality gap, and learn modality-invariant and discriminative features from RGB and IR images.

Existing RGB-infrared person Re-ID approaches mainly concentrate on mitigating the inherent modality discrepancy at the pixel-level or the feature-level to extract cross-modality shared features. For alleviating modality discrepancy at the pixel-level, these methods commonly design complex generative adversarial models (Dai et al. 2018; Wang et al. 2019b, 2020, 2019a) to perform image-to-image translation and generate fake counterpart images, which are difficult to optimize and inevitably introduce noisy generated samples due to the ill-posed infrared-to-RGB transforming. On the other hand, for mitigating modality discrepancy at the feature-level, these methods employ one-stream (Wu et al. 2017; Li et al. 2020) or two-stream networks (Ye et al. 2018a,b, 2020; Hao et al. 2019; Zhu et al. 2020; Ye et al. 2019) to extract modality-invariant features by several customized losses. Nevertheless, one-stream network based methods learn a common network model, which lacks the capacity of explicitly modeling individual modalities and neglects modality specific characteristics, leading to crucial information loss. The two-stream network based methods firstly utilize separate branch layers for each modality to abstract modality-specific informa-

[*] These authors contributed equally.
[†] Corresponding Author

tion, and then use non-branched shared layers for projecting the modality-specific features into a common feature space. They totally separate the process of modeling modality-specific and modality-share information, and could damage the vital cross-modality shared semantic during extracting modality-specific features. Furthermore, all aforementioned methods attempt to directly handle such large modality discrepancy and align two modalities, which are sensitive to the parameters and difficult to converge.

In this work, we propose a novel modality-adaptive mixup and invariant decomposition (MID) approach for RGB-infrared person Re-ID towards learning modality-invariant and discriminative representations. MID designs a modality-adaptive mixup scheme (MAM) to generate appropriate mixed modality images and to reconcile modality gap at the pixel-level, according to the dynamical appearance and modality discrepancies between different RGB and IR images. MID then employs modality-adaptive convolution decomposition (MACD) to simultaneously counter modality discrepancy and enforce cross-modality shared semantics at the feature-level, towards facilitating cross-modality feature learning. Specifically, as shown in Figure 1, MAM is implemented by an actor-critic agent under a deep reinforcement learning framework. The state of the agent is the intermediate feature maps of RGB-infrared image pair, while the relevant action is the mixup ratio at a training step. With the joint optimization of the actor and critic networks, the suitability of the mixup ratio could be progressively boosted, driven by the supervision signal of reward that measures the relative performance improvement of person Re-ID by utilizing the generated mixed modality samples in the evaluation metrics. The actor-critic agent finally performs a dynamical and local linear interpolation policy between different regions of two modality images in a data-adaptive way, and generates augmented mixed modality samples with identity consistency. The mixed modality images mitigate modality discrepancy at the pixel-level and lead to a more continuous modality-invariant latent space. Moreover, MACD is designed to decompose a regular convolution layer into modality-specific basis layers and a modality-shared coefficient layer for handling the discrepancies among RGB, IR and mixed modalities. The former is in charge of modality intrinsic characteristics, while the latter enforces cross-modality shared decomposition coefficient to capture invariant semantics. ResNet-50 model (He et al. 2016) is equipped with MACD to learn aligned and discriminative features of pedestrians. Extensive experimental results on two datasets have shown the effectiveness of the proposed approach.

Note that the related work XIV (Li et al. 2020) also generates X modality as an auxiliary modality to optimize the model. Nevertheless, XIV produces the limited X modality only from RGB images without considering the characteristics of IR images, failing to generate high-quality mixed modality samples for reducing modality gap and guaranteeing modality-invariant in a more continuous latent space.

The main contributions are summarized as follows: (1) We propose a novel modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification towards learning modality-invariant and discriminative representation. (2) We propose a modality-adaptive mixup scheme to generate suitable mixed modality samples for reconciling RGB and IR modalities at the pixel-level. (3) We design a modality-adaptive convolution decomposition to capture invariant visual semantics and shrink the modality discrepancy at the feature-level.

## Related work

### RGB-Infrared Person Re-ID

Existing RGB-infrared person Re-ID approaches mitigate the modality discrepancy at the pixel-level (Dai et al. 2018; Wang et al. 2020; Fan et al. 2020) or the feature-level (Wu et al. 2017; Li et al. 2020; Ye et al. 2018a,b, 2020; Hao et al. 2019). For alleviating modality discrepancy at the pixel-level, Dai *et al.* (Dai et al. 2018) proposed a cross-modality generative adversarial network to handle the large-scale cross-modality metric learning problem. Wang *et al.* (Wang et al. 2020) proposed to generate cross-modality paired-images and perform both global set-level and fine-grained instance-level alignments. Fan *et al.* (Fan et al. 2020) proposed a modality-transfer generative adversarial network to generate a cross-modality counterpart of a source image in the target modality, for obtaining paired images and producing a general and robust unified feature embedding. Zhang *et al.* (Zhang et al. 2021) proposed a teacher-student GAN model (TS-GAN) to generate the corresponding fake IR images and guide the backbone to learn better feature. For mitigating modality discrepancy at the feature-level, Ye *et al.* (Ye et al. 2018a) proposed a hierarchical cross-modality matching model by jointly optimizing the modality-specific and modality-shared metrics. Ye *et al.* (Ye et al. 2018b) proposed a dual-path network with a bi-directional dual-constrained top-ranking loss to learn discriminative feature representations. Luo *et al.* (Ye et al. 2020) proposed a dynamic dual-attentive aggregation (DDAG) learning method by mining both intra-modality part-level and cross-modality graph-level contextual cues for RGB-infrared person Re-ID. Zhu *et al.* (Zhu et al. 2020) proposed a hetero-center loss to reduce the intra-class cross-modality variations and utilized a two-stream part-pooling-based model to obtain modality-invariant features.

### Domain Mixup

Mixup is a recent data augmentation scheme to regularize deep learning models via *global and random linear interpolations* between pairs of samples and their labels, which plays an important role in domain adaption (Zhang et al. 2017; Verma et al. 2019; Xu et al. 2020; Zhong et al. 2020). For example, Zhang *et al.* (Zhang et al. 2017) proposed a simple data augmentation method termed MixUp, which randomly generates virtual training images by linearly interpolating two images and their corresponding labels. Panfilov *et al.* (Panfilov et al. 2019) investigated two regularization including mixup and adversarial unsupervised domain adaptation to improve the generalization of deep learning based knee cartilage segmentation model. Wu *et al.* (Wu, Inkpen, and El-Roby 2020) proposed a dual mixup regularized learning method for unsupervised domain adaptation,

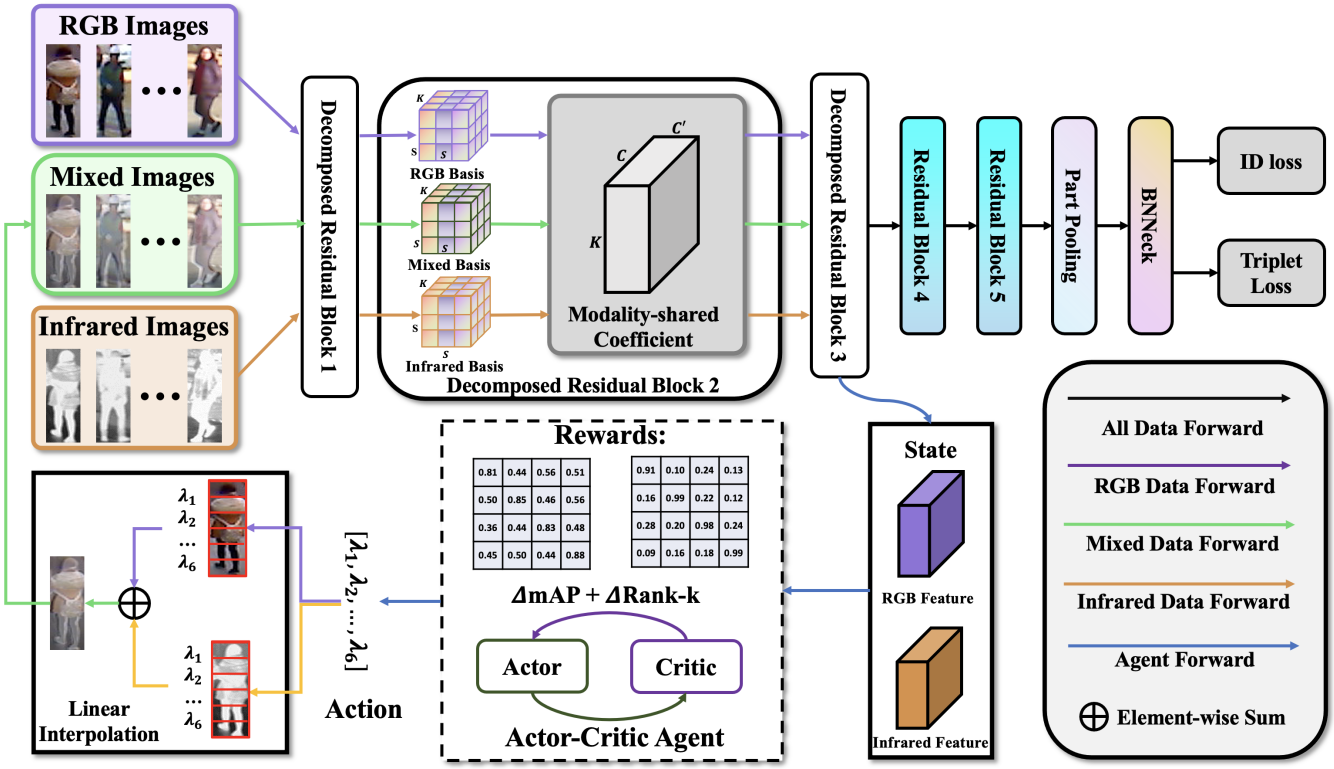


Figure 1: The overall architecture of the proposed MID. It consists of a modality-adaptive mixup module for generating appropriate mixed modality images and reducing the modality discrepancy at the pixel-level, and a decomposed convolution network for further shrinking the modality discrepancy at the feature-level to learn modality-invariant and discriminative representations.

which jointly conducts category and domain mixup regularizations at the pixel-level to enhance the robustness of the model. Xu *et al.* (Xu et al. 2020) presented an adversarial domain adaptation with domain mixup (DM-ADA), which conduct on domain mixup on pixel and feature level to improve the robustness of models. Nevertheless, these existing mixup schemes are not suitable for RGB-Infrared Person Re-ID, which is elaborately analyzed in the next section.

# Method

## Problem Definition and Overview

Let $\boldsymbol{x}^{rgb}$ donates RGB image, and $\boldsymbol{x}^{ir}$ donate IR image. The RGB-IR training dataset is represented as $\boldsymbol{X} = \{\boldsymbol{x}_i^{rgb}, \boldsymbol{x}_i^{ir}, y_i\}_{i=1}^{N}$, where $N$ is the number of pedestrian images. Each image $\boldsymbol{x}^{rgb}$ or $\boldsymbol{x}^{ir}$ corresponds to an identity label $\boldsymbol{Y} = \{y_i\}_{i=1}^{N}$, where $y \in \{1, 2, ..., \mathcal{P}\}$ , and $\mathcal{P}$ refers to the number of persons. The objective is to learn a modality-invariant and discriminative representation for identifying the specific pedestrian from RGB and IR images. We propose a novel modality-adaptive mixup and invariant decomposition (MID) approach for RGB-infrared person Re-ID. As shown in Figure 1, MID firstly employs the modality-adaptive mixup scheme to synthesize augmented mixed modality images between RGB and IR images. The three modality images are then fed into a decomposed convolution network to learn modality-invariant and discriminative features, which is termed as $\mathcal{F}$. The first three residual blocks in $\mathcal{F}$ are denoted as $\mathcal{F}_1$, which are equipped with the designed MACD for absorbing modality variations and aligning invariant semantics across modalities. The decomposed convolution network is optimized with an identification loss and a center triplet loss for re-identification.

## Modality-Adaptive Mixup Scheme

Mixup is an effective data augmentation algorithm that regularizes deep learning models by linear interpolations between pairs of samples and their labels. Nevertheless, existing mixup scheme can not be directly applied to RGB-Infrared Person Re-ID for effectively reconciling RGB and IR modalities at the pixel-level. They commonly generate virtual samples via a simple *global and random* linear interpolation policy. The mixup ratio for interpolating two samples is a scalar, and randomly sampled from a Beta$(\alpha, \alpha)$ distribution with a hyper-parameter $\alpha$. As the inter-modality discrepancy between different RGB and IR images dramatically changes, a random mixup ratio could produce low-quality mixed modality images, which is non-adjacent to real RGB and IR images of the same identity, resulting in manifold intrusion issue. Moreover, as RGB-Infrared Re-ID datasets lack paired RGB and IR images of the same pedestrian and have large intra-modality variations, a scalar mix ratio for globally interpolating RGB and IR images could

corrupt the visual semantic and identity information of the mixed modality image, leading to performance degradation.

For generating appropriate mixed modality images and guaranteeing the identity invariance, we propose a modality-adaptive mixup scheme to bridge the modality discrepancy at the pixel-level. It learns a dynamical and local linear interpolation between the different regions of cross-modality images in data-dependent fashion, which is formulated as Markov decision process and implemented by an actor-critic agent under a deep reinforcement learning (RL) framework (Lillicrap et al. 2016; Li and Chen 2020). Specifically, the modality-adaptive mixup is computed as follows:

$$\begin{aligned} \boldsymbol{x}_{i,g}^{mix} &= m_{i,g}\boldsymbol{x}_{i,g}^{rgb} + (1-m_{i,g})\boldsymbol{x}_{i,g}^{ir}, \\ y_i^{mix} &= y_i^{rgb} = y_i^{ir}, \end{aligned} \tag{1}$$

where $m_{i,g} \in [0,1]$ is the mixup ratio learned from the actor-critic agent with RGB image $\boldsymbol{x}_i^{rgb}$ and IR image $\boldsymbol{x}_i^{ir}$. $\boldsymbol{x}_i^{rgb}$ and $\boldsymbol{x}_i^{ir}$ are divided into $G$ local regions of $\{\boldsymbol{x}_{i,g}^{rgb}\}_{g=1}^{G}$ and $\{\boldsymbol{x}_{i,g}^{ir}\}_{g=1}^{G}$ along horizontal axis, respectively. $\boldsymbol{x}_{i,g}^{mix}$ is the $g$-th local region of the generated mixed modality image. $y_i^{mix}$, $y_i^{rgb}$, $y_i^{ir}$ share the same identity.

The mixup ratio dynamically adjusts based on the modality and appearance discrepancies between the corresponding local regions of RGB and IR images, which is performed by the actor-critic agent. The actor network $\mathcal{A}$ in the agent is employed to estimate the mixup ratio $\boldsymbol{m}_i$, and the critic network $\mathcal{Q}$ in the agent predicts the state-action value (Q-value). $\mathcal{A}$ is formulated as follows:

$$\boldsymbol{m}_i = \sigma\left(\boldsymbol{W}_1\delta(\boldsymbol{W}_0(\mathrm{Pool}(\mathrm{Conv}([\boldsymbol{F}_i^{rgb}, \boldsymbol{F}_i^{ir}]))))\right), \tag{2}$$

where $\boldsymbol{F}_i^{rgb}$ denotes the intermediate feature map of $\boldsymbol{x}_i^{rgb}$ extracted from $\mathcal{F}_1$. $\sigma$ and $\delta$ denote the sigmoid function and rectified linear unit (ReLU) activation function, respectively. $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$ denote two Fully Connected (FC) layers. $\mathrm{Conv}$ denotes a $3\times3$ convolution layer, followed with a batch normalization layer and a ReLU layer. $\mathrm{Pool}$ indicates a global average pooling layer. The critic network in the actor-critic agent has a similar network architecture to the actor network. For guiding the critic network to predict the reliability of the mixup ratio, a reward $\mathcal{R}$ is designed as the supervision signal for optimization. The action, state and reward for the agent are defined as follows:

**State.** The concatenation of the intermediate feature maps $[\boldsymbol{F}_i^{rgb}; \boldsymbol{F}_i^{ir}]$ of the RGB and IR images extracted from $\mathcal{F}_1$ is viewed as the state of the agent.

**Action.** The action of the agent is the mix ratio $\boldsymbol{m}_i \in \mathbb{R}^G$ used for linear interpolation bewteen RGB and IR images. It is a continuous vector.

**Reward.** For a mini-batch input data $\{\boldsymbol{x}_i^{rgb}, \boldsymbol{x}_i^{ir}\}_{i=1}^{b}$ with $y_i^{rgb} = y_i^{ir}$, we can obtain the mixed modality images $\{\boldsymbol{x}_i^{mix}\}_{i=1}^{b}$ via dynamic and local linear interpolation in Eq. (1). We then calculate the similarity matrix $[\boldsymbol{\mathcal{S}}^{rgb,ir}]_{i,j} = \langle \boldsymbol{f}_i^{rgb}, \boldsymbol{f}_j^{ir} \rangle$ between all RGB and IR images, and $[\boldsymbol{\mathcal{S}}^{mix,ir}]_{i,j} = \langle \boldsymbol{f}_i^{mix}, \boldsymbol{f}_j^{ir} \rangle$ between all mixed modality and IR images, where $\boldsymbol{f}_i$ denotes the learned pedestrian representation from the decomposed convolution network $\mathcal{F}$. $\langle \cdot, \cdot \rangle$ denotes Cosine Distance. Thus, the reward is defined as the relative performance improvement of re-identification by using the mixed modality images to enhance the similarity matrix, which is formulated as follows:

$$\begin{aligned} \mathcal{R} =& \mathcal{E}(\boldsymbol{\mathcal{S}}^{rgb,ir} + \boldsymbol{\mathcal{S}}^{mix,ir}) - \mathcal{E}(\boldsymbol{\mathcal{S}}^{rgb,ir}) \\ &+ \mathcal{E}(\boldsymbol{\mathcal{S}}^{ir,rgb} + \boldsymbol{\mathcal{S}}^{mix,rgb}) - \mathcal{E}(\boldsymbol{\mathcal{S}}^{ir,rgb}), \end{aligned} \tag{3}$$

$$\mathcal{E}(\boldsymbol{\mathcal{S}}) = \mathrm{mAP}(\boldsymbol{\mathcal{S}}) + \sum_{k=1}^{K}\frac{1}{k}\,\text{rank-k}(\boldsymbol{\mathcal{S}}), \tag{4}$$

where $\mathrm{mAP}$ and $\text{rank-k}$ are the common evaluation metrics for RGB-Infrared Person Re-ID based on the similarity matrix (Ye et al. 2018a,b). $K$ denotes the number of the pedestrian images for one identity in each modality. We adopt the losses $\mathcal{L}_{\mathcal{A}}$, $\mathcal{L}_{\mathcal{Q}}$ to optimize the actor and the critic networks, respectively, which will be introduced in the following subsection. The data-dependent modality-adaptive mixup scheme explicitly exploits the data distribution of two modalities, and can learn appropriate interpolation policy to generate high-quality mixed modality images for reducing modality discrepancy at the pixel-level and facilitating a more continuous modality-invariant latent space.

## Modality-Adaptive Conv Decomposition

To bridge modality gap at the feature-level, previous works employs one-stream network or two-stream network architecture. Nevertheless, one-stream network based methods lack the capacity of explicitly modeling modality-specific characteristics, resulting in unnecessary information loss. Two-stream network based methods separate the process of modeling modality-specific and modality-share information, vitiating the significant cross-modality shared semantic during extracting modality-specific features. They also require more training parameters.

To address these issues, we propose modality-adaptive convolution decomposition and design a decomposed convolution network $\mathcal{F}$. The decomposed convolution network is built on ResNet-50 model. The modality-adaptive convolution decomposition approximates convolution filters as a linear combination of a small set of dictionary bases, for simultaneously countering modality discrepancy and enforcing cross-modality shared semantic at the feature-level. It is applied to the convolution layers of first three residual blocks $\mathcal{F}_1$ in the decomposed convolution network, each of which is decomposed into modality-specific basis layers and a modality-shared coefficient layer, while both of them remain convolutional. Specifically, $\boldsymbol{W}$ denotes the convolution filters $S \times S \times C_{in} \times C_{out}$ of a convolution layer, where $S$ is the spatial size of the filters, $C_{in}$ and $C_{out}$ denotes the number of input channel and output channel. We decompose $\boldsymbol{W}$ into the modality-specific dictionary bases $\boldsymbol{\alpha} \in \mathbb{R}^{S\times S\times K}$ and the common coefficient across modalities $\boldsymbol{\Psi} \in \mathbb{R}^{K\times C_{in}\times C_{out}}$. In detail, each decomposed convolution layer has three independent modality-specific dictionary bases $\boldsymbol{\alpha}^{rgb}, \boldsymbol{\alpha}^{ir}, \boldsymbol{\alpha}^{mix}$, and the common coefficient $\boldsymbol{\Psi}$ across modalities:

$$\boldsymbol{W}^* = [\boldsymbol{\alpha}^{\{rgb\}}\boldsymbol{\Psi}, \boldsymbol{\alpha}^{\{ir\}}\boldsymbol{\Psi}, \boldsymbol{\alpha}^{\{mix\}}\boldsymbol{\Psi}], \tag{5}$$

The modality-specific dictionary bases are independently learned from the corresponding modality images to model the modality variations. They convolve spatially each individual input feature channel $[\boldsymbol{F}]_{1:C_{in}} \in \mathbb{R}^{H\times W}$ for modality discrepancy correction. The common coefficient is learned from all the three modality images, and performs $1 \times 1$ convolution for weighting sum the corrected output feature channels, thus promoting cross-modality shared semantic. The decomposed convolution network takes RGB, IR and mixed modality images from the modality-adaptive mixup module as input, and effectively handles the large modality gap at the feature-level for learning modality-invariant features. The network outputs global and local features $\boldsymbol{f}_i^{rgb}$, $\boldsymbol{f}_i^{ir}$, $\boldsymbol{f}_i^{mix}$ of three modalities by part pooling operation (Sun et al. 2018).

## Loss Function and Optimization

We adopt the identification loss $L_{id}$ with label smoothing regularization (Ainam et al. 2019) and the center triplet loss $L_{ct}$ (He et al. 2018; Zhao et al. 2020) to optimize the decomposed convolution network for re-identification. The identification loss is calculated as follows:

$$L_{id} = \sum_{i=1}^{\mathcal{P}} -q_i \log\left(p_i\right),\ q_i = \begin{cases} 1 - \frac{\mathcal{P}-1}{\mathcal{P}}\xi, & y = i \\ \frac{\xi}{\mathcal{P}}, & y \neq i \end{cases} \quad (6)$$

where $\mathcal{P}$ is the number of identities in the training set, $y$ is the ground-truth ID and $p_i$ denotes the ID prediction logits of the $i^{th}$ class. $\xi$ is the smoothing parameter, which is beneficial to prevent the network from over-fitting to training IDs. The three identification losses $L_{id}^{rgb}$ , $L_{id}^{ir}$ , $L_{id}^{mix}$ are used to supervise the global and local features of three modalities $\boldsymbol{f}_i^{rgb}$, $\boldsymbol{f}_i^{ir}$, $\boldsymbol{f}_i^{mix}$, respectively. For the center triplet loss, we randomly sample $P$ identities and $K$ images of each identity for RBG and IR modality to form a mini-batch with $PK$ images termed $\{\boldsymbol{x}_{p,k}^{rgb}, \boldsymbol{x}_{p,k}^{ir}, y_p\}_{p=1,k=1}^{P,K}$. We calculate the feature centers of each pedestrian for the three modality images in a mini-batch, $\boldsymbol{c}_p^{\{rgb,ir,mix\}} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{f}_k^{\{rgb,ir,mix\}}$. Thus, the center triplet loss is defined as follows:

$$\begin{aligned} L_{ct}^{\alpha,\beta} =& \sum_{i=1}^{P}[\rho + \|c_p^{\alpha} - \boldsymbol{c}_p^{\beta}\|_2 - \min_{\substack{n\in\{\alpha,\beta\}\\ j\neq p}} \|\boldsymbol{c}_p^{\alpha} - \boldsymbol{c}_j^{n}\|_2]_+ \\ &+ \sum_{i=1}^{P}[\rho + \|\boldsymbol{c}_p^{\beta} - \boldsymbol{c}_p^{\alpha}\|_2 - \min_{\substack{n\in\{\alpha,\beta\}\\ j\neq p}} \|\boldsymbol{c}_p^{\beta} - \boldsymbol{c}_j^{n}\|_2]_+, \end{aligned} \quad (7)$$

where $\rho$ is a margin parameter, $[z]_+ = max(z, 0)$, $\alpha, \beta \in \{rgb, ir, mix\}$ which denote two different modalities. We utilize three center triplet losses $L_{ct}^{rgb,ir}, L_{ct}^{rgb,mix}, L_{ct}^{ir,mix}$ to supervise the three modality features $\boldsymbol{f}_i^{rgb}$, $\boldsymbol{f}_i^{ir}$, $\boldsymbol{f}_i^{mix}$. Therefore, the total loss $L_{dcn}$ for the decomposed convolution network is the combination of these identification and center triplet losses:

$$\begin{aligned} L_{dcn} =& \lambda_1 L_{ct}^{rgb,ir} + \lambda_2 L_{ct}^{rgb,mix} + \lambda_3 L_{ct}^{ir,mix} \\ &+ \lambda_4 L_{id}^{rgb} + \lambda_5 L_{id}^{ir} + \lambda_6 L_{id}^{mix}, \end{aligned} \quad (8)$$

where $\lambda_{1-6}$ is the trade-off parameters.

Different from standard reinforcement learning algorithm, the proposed actor-critic agent does not have explicit sequential relationship along different training steps. The action of the interpolation policy between RGB and IR images is conditioned on the state of their intermediate features in a one-shot fashion, which is essentially a one-step Markov decision process. The action space is continuous, thus the optimal action could be found by gradient ascent method following the solution of continuous Q-value prediction (Lillicrap et al. 2016; Li and Chen 2020). The loss for actor network is defined as follows:

$$L_{\mathcal{A}} = -\mathcal{Q}(\mathcal{A}([\boldsymbol{F}^{rgb}, \boldsymbol{F}^{ir}]), [\boldsymbol{F}^{rgb}, \boldsymbol{F}^{ir}]), \quad (9)$$

The actor network $\mathcal{A}$ is updated to achieve higher Q-value, which implies higher rank-$k$ accuracy and mAP for person Re-ID. The critic network is optimized to predict an accurate Q-value estimation, thus MSE loss is employed as follows:

$$L_{\mathcal{Q}} = \|\mathcal{Q}(\mathcal{A}([\boldsymbol{F}^{rgb}, \boldsymbol{F}^{ir}]), [\boldsymbol{F}^{rgb}, \boldsymbol{F}^{ir}]) - \mathcal{R}\|^2, \quad (10)$$

For the entire MID, we combine the advantage of supervised and reinforcement learning, alternately optimizing the decomposed convolution network and the actor-critic agent.

# Experiments

## Experimental Setting

**Datasets.** We evaluate the proposed MID using two public RGB-Infrared datasets: RegDB (Nguyen et al. 2017) and SYSU-MM01 (Wu et al. 2017). RegDB dataset contains 412 pedestrians. Each pedestrian has 10 visible images and 10 thermal images. Following the evaluation protocol (Ye et al. 2018a,b), this dataset is randomly split into two parts, 206 identities for training and the other 206 identities for testing, with two different testing modes, *i.e.*, visible to thermal mode and thermal to visible mode. The reported results are the average of 10 random training/test splits on RegDB dataset. SYSU-MM01 (Wu et al. 2017) is the largest existing RGB-infrared dataset, which was captured with 4 visible and 2 infrared cameras. The training set contains 395 persons with 22,258 RGB images and 11,909 IR images, while the testing set contains 96 persons with 3,803 IR images and 301 RGB images. We adopt the all-search mode and indoor-search mode (Wu et al. 2017; Luo et al. 2019) to evaluate the performance.

**Evaluation Metrics.** The standard Cumulative Matching Characteristic (CMC) at Rank-$k$ and the mean Average Precision (mAP) are adopted as evaluation metrics.

**Implementation Details:** The proposed method is implemented by the PyTorch framework with one NVIDIA Tesla V100 GPU. Each mini-batch contains 96 images of 8 identities (each person has 4 RGB images, 4 IR images, and 4 generated mixed modality images). ResNet-50 (He et al. 2016) model is adopted as the backbone network. Part-pooling (Sun et al. 2018) is added after the backbone. The first three residual blocks of ResNet-50 model are equipped with modality-adaptive convolution decomposition. The stride of the last convolution layer is set to 1. The margin $\rho$ is set to 0.3. The parameter $\mu$ and $\xi$ are set to 1

Table 1: Performance (%) comparison to the state-of-the-art methods on RegDB and SYSU-MM01 datasets.

| Methods | RegDB | | | | | | SYSU-MM01 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Visible to Thermal | | | Thermal to Visible | | | All | | | Indoor | | |
| | r1 | r10 | mAP | r1 | r10 | mAP | r1 | r10 | mAP | r1 | r10 | mAP |
| Zero-Pad (Wu et al. 2017) | 17.75 | 34.21 | 18.90 | 16.63 | 34.68 | 17.82 | 14.80 | 54.12 | 15.95 | 20.58 | 68.38 | 26.92 |
| HCML (Ye et al. 2018a) | 24.44 | 47.53 | 20.80 | 21.70 | 45.02 | 22.24 | 14.32 | 53.16 | 16.16 | 24.52 | 73.25 | 30.08 |
| MAC (Ye, Lan, and Leng 2019) | 36.43 | 62.36 | 37.03 | 36.20 | 61.68 | 39.23 | 33.26 | 79.04 | 36.22 | 36.43 | 62.36 | 37.03 |
| MSR (Feng, Lai, and Xie 2019) | 48.43 | 70.32 | 48.67 | - | - | - | 37.35 | 83.40 | 38.11 | 39.64 | 89.29 | 50.88 |
| D-HSME (Hao et al. 2019) | 50.85 | 73.36 | 47.00 | 50.15 | 72.40 | 46.16 | 20.68 | 62.74 | 23.12 | - | - | - |
| EDFL (Liu et al. 2020) | 52.58 | 72.10 | 52.98 | 51.89 | 72.09 | 52.13 | 36.94 | 85.42 | 40.77 | - | - | - |
| AlignGAN (Wang et al. 2019a) | 57.90 | - | 53.60 | 56.30 | - | 53.40 | 42.40 | 85.00 | 40.70 | 45.90 | 87.60 | 54.30 |
| TS-GAN (Zhang et al. 2021) | - | - | - | - | - | - | 49.8 | 87.3 | 47.4 | 50.4 | 90.8 | 63.1 |
| XIV (Li et al. 2020) | 62.21 | 83.13 | 60.18 | - | - | - | 49.92 | 89.79 | 50.73 | - | - | - |
| DDAG (Ye et al. 2020), | 69.34 | 86.19 | 63.46 | 68.06 | 85.15 | 61.80 | 54.75 | 90.39 | 53.02 | 61.02 | 94.06 | 67.98 |
| Hi-CMD (Choi et al. 2020) | 70.93 | 86.39 | 66.04 | - | - | - | 34.94 | 77.58 | 35.94 | - | - | - |
| TSLFN (Zhu et al. 2020) | - | - | - | - | - | - | 56.96 | 91.50 | 54.95 | 59.74 | 92.07 | 64.91 |
| HAT (Ye, Shen, and Shao 2020) | 71.83 | 87.16 | 67.56 | 70.02 | 86.45 | 66.30 | 55.29 | 92.14 | 53.89 | 62.10 | 95.75 | 69.37 |
| $G^2$DA (Wan et al. 2021) | 71.72 | 87.13 | 65.90 | 69.50 | 84.87 | 63.88 | 57.07 | 90.99 | 55.05 | 63.70 | 94.06 | 69.83 |
| NFS (Chen et al. 2021) | 80.54 | 91.96 | 72.10 | 77.95 | 90.45 | 69.79 | 56.91 | 91.34 | 55.45 | 62.79 | **96.53** | 69.79 |
| MID (ours) | **87.45** | **95.73** | **84.85** | **84.29** | **93.44** | **81.41** | **60.27** | **92.90** | **59.40** | **64.86** | 96.12 | **70.12** |

and 0.1, respectively. The trade-off parameters $\lambda_{1,4,5}$ are set to 1, $\lambda_{2,3}$ are set to 0.5, and $\lambda_6$ is set to 0.1 in Eq. (8). We adopt Adam Optimizer to train the actor-critic agent. And we utilize the stochastic gradient descent (SGD) optimizer for MACD with the momentum of 0.9, the initial learning rate of 0.05, 0.02 on RegDB and SYSU-MM01 datasets, respectively. The learning rates decayed by 0.1 after 20 and 45 epochs. The whole MID framework is trained for 60 epochs on RegDB dataset which takes 1 hour, and for 100 epochs on SYSU-MM01 dataset which takes 6 hours.

## Comparison to State-of-the-Art Methods

**RegDB:** We present a comparison the proposed MID with 11 state-of-the-art approaches on RegDB dataset in Table 1. Experiments conducted on RegDB dataset show that the proposed MID achieves the best performance under different testing modes over all state-of-the-art methods by large margins. For Visible to Thermal mode, MID achieves 87.45% rank-1 accuracy and 84.85% mAP, outperforming the 2nd best NFS (Chen et al. 2021) by 6.91% rank-1 accuracy and 12.75% mAP, respectively. For Infrared to RGB mode, MID also obtains 84.29% rank-1 accuracy and 84.29% mAP, improving the 2nd best method NFS (Chen et al. 2021) by 6.34% rank-1 accuracy and 11.62% mAP, respectively. The boosting demonstrates that the proposed MID is able to learn modality-invariant and discriminative features from cross-modality RGB and infrared images.

**SYSU-MM01:** Table 1 also reports the performance of the proposed MID with 11 state-of-the-art methods on SYSU-MM01 dataset. Experiments results show that MID obtains the best performance under both All-search and Indoor settings. For all-search mode, MID achieves 60.27% rank-1 accuracy and 59.40% mAP, outperforming the 2nd best method NFS (Wan et al. 2021) by 3.20% rank-1 accuracy and 4.35% mAP score. For indoor-search mode, MID also obtains the best rank-1 accuracy and mAP score. The comparisons demonstrate that MID can effectively reduce inter-modality and intra-modality discrepancies.

Table 2: Ablation studies on the effectiveness of each component of the proposed MID.

| $\mathcal{B}$ | $\mathcal{M}$ | $\mathcal{D}$ | RegDB | | SYSU-MM01 | |
|---|---|---|---|---|---|---|
| | | | r1 | mAP | r1 | mAP |
| ✓ | × | × | 76.34 | 70.81 | 49.46 | 49.32 |
| ✓ | ✓ | × | 83.43 | 79.13 | 56.22 | 57.12 |
| ✓ | ✓ | ✓ | 87.45 | 84.85 | 60.27 | 59.40 |

## Ablation Study

**Effectiveness of each component of MID.** We conduct ablation studies on RegDB and SYSU-MM01 datasets to investigate the effectiveness of the components of MID in Table 2, including modality-adaptive mixup scheme ($\mathcal{M}$) and modality-adaptive convolution decomposition ($\mathcal{D}$). $\mathcal{B}$ denotes a vanilla ResNet-50 model trained with identity loss and normal triplet loss. When introducing the modality-adaptive mixup scheme, the performance is improved by 7.09% rank-1 accuracy and 8.32% mAP on RegDB dataset, and by 6.76% rank-1 accuracy and 7.80% mAP on SYSU-MM01 dataset. The boosting indicates the dynamical and local linear interpolation policy of $\mathcal{M}$ can generate appropriate mixed modality images and mitigate the inherent modality discrepancy at the pixel-level for promoting a more continuous modality-invariant latent space. By adding the modality-adaptive convolution decomposition, the performance is improved by 4.02% rank-1 accuracy and 5.72% mAP on RegDB dataset, and by 4.05% rank-1 accuracy and 2.28% mAP on SYSU-MM01 dataset. This demonstrates that the decomposed convolution network with $\mathcal{D}$ can capture invariant visual semantics and further shrink the modality discrepancy at the feature-level.

**Analysis of modality-adaptive mixup scheme.** Table 3 shows the influence of different mixup schemes. Fix refers

Table 3: Influence of different mixup schemes for MID.

| Mixup Scheme | RegDB | | SYSU-MM01 | |
|---|---|---|---|---|
| | r1 | mAP | r1 | mAP |
| Fix | 83.96 | 77.83 | 50.43 | 48.64 |
| Beta | 83.53 | 78.04 | 51.46 | 50.09 |
| MAM | 87.45 | 84.85 | 60.27 | 59.40 |

Table 4: Influence of different number $G$ of the partitioned local regions for mixup.

| Mixup Scheme | RegDB | | SYSU-MM01 | |
|---|---|---|---|---|
| | r1 | mAP | r1 | mAP |
| w/o partition | 83.93 | 79.04 | 53.46 | 52.09 |
| 5 | 86.22 | 82.53 | 58.22 | 59.33 |
| 6 | 87.45 | 84.85 | 60.27 | 59.40 |
| 7 | 85.45 | 81.79 | 57.33 | 56.27 |

to a fixed mixup ratio of 0.5 for interpolating RGB and IR images, Beta refers to a mixup ratio randomly sampled from Beta prior distribution. MAM refers to our modality-adaptive mixup scheme. Fix and Beta belong to data-independent mixup schemes, while MAM is data-dependent. We can observe that the data-independent mixup schemes Fix and Beta obtain performance degradation over the data-dependent modality-adaptive mixup scheme. The comparison demonstrates that the proposed MAM can learn the dynamical modality discrepancy between different pair of RGB and IR images, and adaptively adjust the mixup ratio to generate appropriate mixed modality images for promoting a more continuous modality-invariant latent space and extracting more effective RGB-IR representations.

The results in Table 4 show the influence of different number $G$ of the partitioned local regions for the modality-adaptive mixup scheme on RegDB dataset. W/o partition refers to the employ of a global mixup ratio to interpolate the whole RGB and IR images. We can observe that local linear interpolation policy ($\boldsymbol{m} \in \mathbb{R}^G$) obtains obvious performance improvement over global linear interpolation policy ($\boldsymbol{m} \in \mathbb{R}^1$), which indicates that adaptively mixing different local regions of RGB and IR images could reduce intra-modality discrepancy, and produce high-quality mixed modality images with identity consistency and fewer confusing information (*e.g.*, occlusion, blurring and ghosting) for promoting cross-modality feature learning. Moreover, Table 4 shows that when adopting 6 partitioned local regions, the proposed MID with $\boldsymbol{m} \in \mathbb{R}^6$ obtains the best re-identification performance.

**Analysis of modality-adaptive convolution decomposition.** Table 5 shows the influence of the number of residual blocks equipped with modality-adaptive convolution decomposition. From Table 5, we can observe that when $n_d$ increases from 0 to 3, the proposed method obtains 4.02% and 4.05% improvements in rank-1 accuracy on RegDB and SYSU-MM01 datasets, respectively. When $n_d$ increases from 3 to 5, the performance decreases by 4.87% and 5.76% rank-1 accuracy on RegDB and SYSU-MM01 datasets. MID obtains the best performance when $n_d = 3$, indicting that

Table 5: Influence of different number of the decomposed residual blocks $n_d$.

| Mixup Scheme | RegDB | | SYSU-MM01 | |
|---|---|---|---|---|
| | r1 | mAP | r1 | mAP |
| 0 | 83.43 | 79.13 | 56.22 | 57.12 |
| 1 | 85.94 | 82.69 | 56.71 | 58.45 |
| 2 | 86.71 | 83.54 | 57.52 | 58.49 |
| 3 | 87.45 | 84.85 | 60.27 | 59.40 |
| 4 | 84.38 | 82.27 | 56.75 | 57.25 |
| 5 | 82.58 | 80.99 | 54.51 | 56.58 |

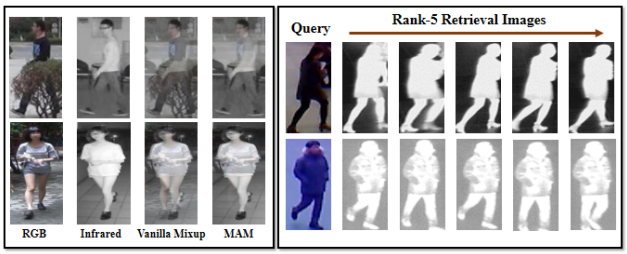


Figure 2: (a) Visualization of the generated mixed modality images by the vanilla mixup scheme and the modality-adaptive mixup scheme; (b) Visualization of some retrieval results on RegDB dataset.

suitable decomposed residual blocks can capture more invariant visual semantics and improve the capacity of the learned modality-shared representations.

**Visualization of results.** To further verify the benefit of the modality adaptive mixup scheme, we visualize some mixed modality images in Figure 2(a). Compared with the vanilla mixup scheme (Zhang et al. 2017), the modality adaptive mixup scheme is able to generate high-quality mixed modality images without the interference of occlusion, blurring and ghosting, *etc*. Moreover, we visualize some retrieval results in Figure 2(b), which demonstrate the effectiveness of the proposed MID for identifying the same pedestrians and distinguishing different pedestrians.

## Conclusions

In this work, we propose a novel modality-adaptive mixup and invariant decomposition (MID) approach for RGB-infrared person re-identification to mitigate the inherent modality discrepancy between RGB and IR images at the pixel-level and feature-level. MID firstly introduces modality-adaptive mixup scheme to generate appropriate mixed modality images by the actor-critic agent for reducing modality gap at the pixel-level and facilitating a more continuous modality-invariant latent space. MID then designs modality-adaptive convolution decomposition to simultaneously counter modality discrepancy and enforce cross-domain shared semantics at the feature-level, for learning effective modality-shared representation. Experimental results on two cross-modality person re-identification datasets have demonstrated the superiority of the proposed method.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62106245, and the Fundamental Research Funds for the Central Universities under Grant WK2100000021.

## References

Aggarwal, J. K.; and Ryoo, M. S. 2011. Human activity analysis: A review. *CSUR*.

Ainam, J.-P.; Qin, K.; Liu, G.; and Luo, G. 2019. Sparse label smoothing regularization for person re-identification. *IEEE Access*.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, Y.; Wan, L.; Li, Z.; Jing, Q.; and Sun, Z. 2021. Neural Feature Search for RGB-Infrared Person Re-Identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Dai, P.; Ji, R.; Wang, H.; Wu, Q.; and Huang, Y. 2018. Cross-modality person re-identification with generative adversarial training. In *Proceedings of the International Joint Conferences on Artificial Intelligence*.

Fan, X.; Jiang, W.; Luo, H.; and Mao, W. 2020. Modality-transfer generative adversarial network and dual-level unified latent representation for visible thermal Person re-identification. *The Visual Computer*.

Feng, Z.; Lai, J.; and Xie, X. 2019. Learning modality-specific representations for visible-infrared person re-identification. *IEEE Transactions on Image Processing*.

Hao, Y.; Wang, N.; Li, J.; and Gao, X. 2019. HSME: Hypersphere manifold embedding for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

He, X.; Zhou, Y.; Zhou, Z.; Bai, S.; and Bai, X. 2018. Triplet-center loss for multi-view 3d object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Kim, Y.-H.; Jeon, J. H.; Lee, B.; Choe, E. K.; and Seo, J. 2017. OmniTrack: A flexible self-tracking approach leveraging semi-automated tracking. *IMWUT*.

Li, D.; and Chen, Q. 2020. Deep Reinforced Attention Learning for Quality-Aware Visual Recognition. In *Proceedings of the European Conference on Computer Vision*.

Li, D.; Wei, X.; Hong, X.; and Gong, Y. 2020. Infrared-Visible Cross-Modal Person Re-Identification with an X Modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2016. Continuous control with deep reinforcement learning. In *Proceedings of the International Conference on Learning Representations*.

Liu, H.; Cheng, J.; Wang, W.; Su, Y.; and Bai, H. 2020. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. *Neurocomputing*.

Luo, C.; Chen, Y.; Wang, N.; and Zhang, Z. 2019. Spectral feature transformation for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*.

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. *IEEE Sensors Journal*.

Panfilov, E.; Tiulpin, A.; Klein, S.; Nieminen, M. T.; and Saarakkala, S. 2019. Improving robustness of deep learning based knee mri segmentation: Mixup and adversarial domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

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 *Proceedings of the European Conference on Computer Vision*.

Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold mixup: Better representations by interpolating hidden states. In *ICML*.

Wan, L.; Sun, Z.; Jing, Q.; Chen, Y.; Lu, L.; and Li, Z. 2021. $G^2$DA: Geometry-Guided Dual-Alignment Learning for RGB-Infrared Person Re-Identification. *arXiv preprint arXiv:2106.07853*.

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 *Proceedings of the IEEE International Conference on Computer Vision*.

Wang, G.-A.; Yang, T. Z.; Cheng, J.; Chang, J.; Liang, X.; and Hou, Z. 2020. Cross-Modality Paired-Images Generation for RGB-Infrared Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wang, X. 2013. Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*.

Wu, Y.; Inkpen, D.; and El-Roby, A. 2020. Dual mixup regularized learning for adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision*.

Xu, M.; Zhang, J.; Ni, B.; Li, T.; Wang, C.; Tian, Q.; and Zhang, W. 2020. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ye, M.; Lan, X.; and Leng, Q. 2019. Modality-aware collaborative learning for visible thermal person re-identification. In *Proceedings of the ACM International Conference on Multimedia*.

Ye, M.; Lan, X.; Li, J.; and Yuen, P. 2018a. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ye, M.; Lan, X.; Wang, Z.; and Yuen, P. C. 2019. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*.

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*.

Ye, M.; Sheng, J.; Crandall, D. J.; Shao, L.; and Luo, J. 2020. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Proceedings of the European Conference on Computer Vision*.

Ye, M.; Wang, Z.; Lan, X.; and Yuen, P. C. 2018b. Visible thermal person re-identification via dual-constrained top-ranking. In *Proceedings of the International Joint Conferences on Artificial Intelligence*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations*.

Zhang, Z.; Jiang, S.; Huang, C.; Li, Y.; and Da Xu, R. Y. 2021. RGB-IR cross-modality person ReID based on teacher-student GAN model. *Pattern Recognition Letters*.

Zhao, C.; Lv, X.; Zhang, Z.; Zuo, W.; Wu, J.; and Miao, D. 2020. Deep Fusion Feature Representation Learning With Hard Mining Center-Triplet Loss for Person Re-Identification. *IEEE Trans Multimedia*.

Zhong, Z.; Zhu, L.; Luo, Z.; Li, S.; Yang, Y.; and Sebe, N. 2020. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. *arXiv preprint arXiv:2004.05551*.

Zhu, Y.; Yang, Z.; Wang, L.; Zhao, S.; Hu, X.; and Tao, D. 2020. Hetero-center loss for cross-modality person re-identification. *Neurocomputing*.