Infrared-Visible Cross-Modal Person Re-Identification with an *X* Modality

Diangang Li,[1] **Xing Wei,**[1*] **Xiaopeng Hong,**[1,3] **Yihong Gong**[2]
[1]Faculty of Electronic and Information Engineering, Xi'an Jiaotong University
[2]School of Software Engineering, Xi'an Jiaotong University
[3]Research Center for Artificial Intelligence, Peng Cheng Laboratory
{diangangli, xingxjtu}@gmail.com, {hongxiaopeng, ygong}@mail.xjtu.edu.cn

Abstract

This paper focuses on the emerging Infrared-Visible cross-modal person re-identification task (IV-ReID), which takes infrared images as input and matches with visible color images. IV-ReID is important yet challenging, as there is a significant gap between the visible and infrared images. To reduce this 'gap', we introduce an auxiliary *X* modality as an assistant and reformulate infrared-visible *dual-mode* cross-modal learning as an X-Infrared-Visible *three-mode* learning problem. The *X* modality restates from RGB channels to a format with which cross-modal learning can be easily performed. With this idea, we propose an X-Infrared-Visible (XIV) ReID cross-modal learning framework. Firstly, the *X* modality is generated by a lightweight network, which is learnt in a self-supervised manner with the labels inherited from visible images. Secondly, under the XIV framework, cross-modal learning is guided by a carefully designed modality gap constraint, with information exchanged cross the visible, *X*, and infrared modalities. Extensive experiments are performed on two challenging datasets SYSU-MM01 and RegDB to evaluate the proposed XIV-ReID approach. Experimental results show that our method considerably achieves an absolute gain of over 7% in terms of rank 1 and mAP even compared with the latest state-of-the-art methods.

Introduction

Person re-identification (ReID) aims at identifying target persons in a query set from a large-scale gallery set captured by non-overlapping camera views (Zheng, Yang, and Hauptmann 2016; Zhang et al. 2015; Cheng et al. 2016). Its great application value in surveillance has propelled ever-increasing research efforts (Luo et al. 2019; Yang et al. 2019) as many other computer vision tasks (Wang et al. 2010; Ma et al. 2019). Encouraging performance has been observed, especially in the visible spectrum, where all the images are captured by visible cameras (Tay, Roy, and Yap 2019). However, visible cameras cannot provide enough discriminative information under poor lighting conditions, *e.g.*, in the dark. The applications are thus limited if only using visible cameras.

*Corresponding author

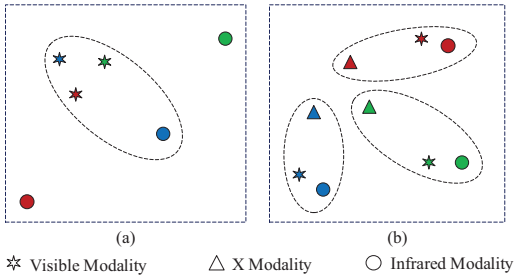


Figure 1: Conceptual illustration of cross-modal learning with '*X*'. Different colors represent different IDs. And images in dashed box are easily to match together in prediction. (a) Cross-model matching is difficult as the gap between visible and infrared modality is clear; (b) With '*X*' (in triangles), cross-modal matching becomes easier. (**Best viewed in color.**)

For practical use, modern surveillance systems usually operate in dual-modes, *i.e.*, working in the visible mode during the day and automatically switching to the infrared mode at night. Consequently, a new task arises naturally (Wang et al. 2019c). Given a target infrared image, the goal is to match the visible images of the corresponding person. This cross-modal image matching task is named Infrared-Visible person re-identification (IV-ReID).

A few studies have been proposed for this IV-ReID task (Ye et al. 2018a; 2018b; Dai et al. 2018; Wang et al. 2019c; Feng, Lai, and Xie 2019; Hao et al. 2019; Wang et al. 2019a). One special challenge in this task is how to bridge the modality gap between the visible and infrared images. Visible images have three channels containing colour information of visible light with wavelengths from 400 nm to 700 nm, while infrared images have one-channel of invisible electromagnetic radiation, the wavelengths of which are between 700 nm and 1 mm, longer than those of visible light. The two modalities are, thus, inherently different. Such a difference further accounts for different ways of using the two modalities. For visible images based person ReID models, the learned high-level semantic features mainly cover appearance and colour information (Zhong et al. 2019;

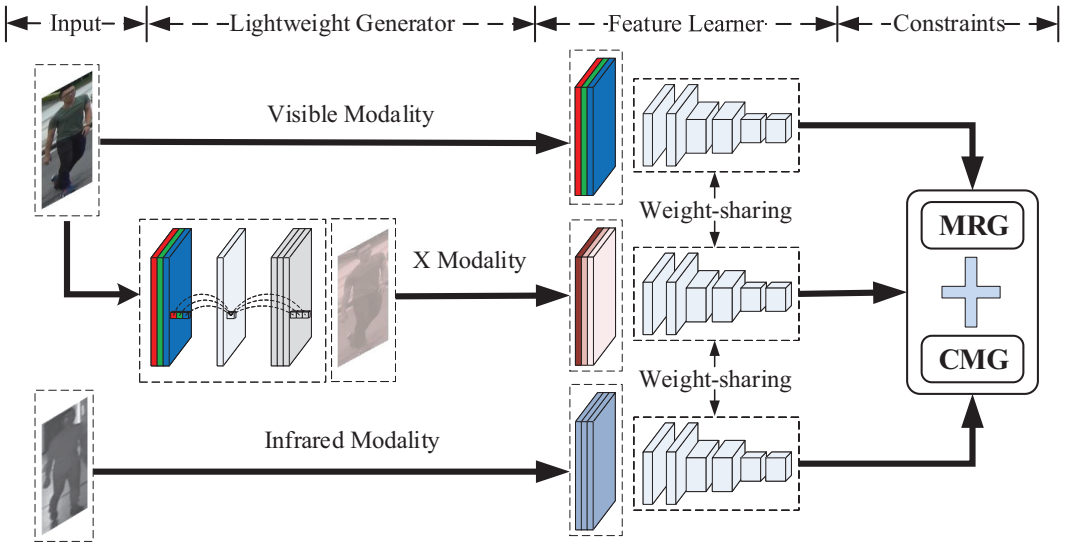


Figure 2: Illustration of the proposed XIV cross-modal learning framework with an adjoint and auxiliary X modality. The lightweight generator siphons off knowledge from the visible and infrared images and outputs the X modality images. Then three modalities are fed into the weight-sharing cross-modal feature learner. And two modality constraints namely the cross modality gap (CMG) and the modality respective gap (MRG) constraint, are designed to regularize feature representations and classification outputs and to learn cross-modal information of three modalities in a common space.

Sun and Zheng 2019). On the other hand, the infrared images based ones mainly include structure and shape information (Jungling and Arens 2010). As a result, when enforcing these two modalities directly into a joint objective function for cross-modal learning, most existing IV-ReID approaches become sensitive to the parameters, difficult to converge, and computationally intensive. It is not surprising that most of the state-of-the-arts only achieve a mean Average Precision under 30% on the challenging dataset SYSU-MM01 (Ye et al. 2018a; 2018b; Hao et al. 2019; Dai et al. 2018; Wang et al. 2019c).

To mitigate the gap problem caused by different modalities, we propose a new infrared-visible cross-modal person ReID approach, termed X-Infrared-Visible (XIV) cross-modal learning. The X modality is an adjoint, auxiliary modality to reconcile both the infrared and visible modalities. With the assistance of X, the visible and infrared modalities are connected, and cross-modal learning becomes easier, as demonstrated by Fig. 1.

The proposed XIV cross-modal learning approach consists of two main components: a lightweight X modality generator and a weight-sharing XIV cross-modal feature learner. Concretely, the lightweight network generates the X modality images with a very small amount of extra costs. It uses visible images as input, siphons off knowledge from the visible and infrared images through self-supervision (Agrawal, Carreira, and Malik 2015; Sun et al. 2019), and outputs the X images. Afterwards, the XIV feature learner takes the three modalities, namely the X, the infrared, and the visible ones, as input. Through weight sharing, the feature learner adapts to three modalities jointly and generates modality-invariant features in the common space, where cross-modal person ReID can be efficiently performed. Subsequently, the X modality generator is linked to the feature learner so that the proposed XIV approach can be optimized in an end-to-end manner. We devise a modality gap constraint to direct the knowledge communication across modalities and optimization through back-propagation during learning, by jointly considering the infrared-X cross-modal disparity, infrared-visible cross-modal disparity, and the respective constraints for each modality. Finally, we evaluate the proposed framework on two publicly available datasets SYSU-MM01 (Wu et al. 2017) and RegDB (Nguyen et al. 2017).

The main contributions are summarized as follows:

- We propose a new approach for infrared-visible cross-modal person ReID through the X-Infrared-Visible cross-modal learning, based on an adjoint and auxiliary X modality.
- We propose an extra lightweight network to generate the X modality through self-supervised learning.
- We devise a modality gap constraint to direct the learning and knowledge communication across modalities.
- We advance the latest state-of-the-art performance of rank 1 and mAP accuracy over 7% on the large-scale SYSU-MM01 dataset.

Related Work

Self-supervised learning. As an alternative to fully-supervised algorithms, self-supervised learning has provided considerable improvements in cross-modal learning, like image and context cross-modal learning (Doersch,

Gupta, and Efros 2015; Gomez et al. 2017), instructions and trajectories cross-modal learning (Wang et al. 2019b), *etc*. Cross-modal self-supervised learning algorithms attempt to utilize the structure in one modality to provide the training supervision for co-occurring modality (Patel et al. 2019).

Considering self-supervision in single-modality person ReID, Sun et al. (Sun et al. 2019) propose a visibility-aware part model, which learns to perceive the visibility of regions through self-supervision. Specifically, they randomly crop partial pedestrian images from the holistic ones and automatically generate corresponding labels.

Infrared-visible person re-identification. For the IV-ReID problem, the cross-modal methods mainly try to learn modality relevant features. Wu et al. (Wu et al. 2017) propose a two-stream deep zero-padding network to learn the cross-modal features in a common space. Ye et al. (Ye et al. 2018a) introduce a two-step framework combining both feature learning and metric learning. They (Ye et al. 2018b) also propose a cross-modal pair-wise constraint to narrow the gap between visible features and infrared features. Hao et al. (Hao et al. 2019) propose an end-to-end dual stream hyper-sphere manifold embedding model to constrain the intra-modality and inter-modality variations. And the framework proposed in (Feng, Lai, and Xie 2019) utilizes the modality-related information and extracts modality-specific representations by constructing an individual network for each modality.

The works most relevant to ours are three GAN-based methods, cmGAN (Dai et al. 2018), D^{2}RL (Wang et al. 2019c) and AlignGAN (Wang et al. 2019a). cmGAN adopts generative adversarial networks (Goodfellow et al. 2014) (GANs) to learn discriminative representations from different modalities. D^{2}RL adopts GANs to reduce the modality discrepancy and appearance discrepancy separately. Similarly, AlignGAN accomplishes pixel and feature alignment in an unified GAN framework. However, the differences to ours are evident. Firstly, compared with cmGAN which directly works on the original two modalities with weight-sharing networks, we adopt an auxiliary *X* modality to narrow the modality gap between the original two with the devised modality gap constraints and ease the learning difficulty just with a lightweight network and slight costs. Secondly, in D^{2}RL, GANs are applied to translate the visible (or infrared) image to its infrared (or visible) counterpart. Then the visible (or infrared) image is stacked with its infrared (or visible) counterpart to form a four-channel multi-spectrum image which is fed into a backbone network as input. Comparatively, we generate an individual *X* modality and perform efficient cross-modal learning using three three-channel modalities. Our new XIV cross-modal learning sheme is more advisable since the ill-posed infrared-to-visible generation in D^{2}RL is avoided and most modern person ReID backbones can be directly used without extra workload to change the structures and finetune. Thirdly, in AlignGAN, GANs are adopted to map a visible image to a fake infrared image, demanding that the generated fake infrared image could be mapped back to original visible one. At first, as infrared image contains much less information than visible image, the (fake) infrared-to-visible ill-posed translation still remains, especially with the spatial transformation in GANs. Instead of focusing on mapping visible images to infrared images directly, we intend to learn an intermediate mediator between visible and infrared images. Generally, the GANs used in these methods are complicated and difficult to train. In contrast, we utilize an extremely lightweight network as the *X* modality generator, which is much easier to optimize than GANs. In experiments, we will show that our method greatly outperforms those methods.

Methodology

Problem Formulation

We denote the total cross-modal person ReID dataset as $\mathcal{T} = \{\mathcal{T}_{tr}, \mathcal{T}_{te}\}$. Suppose that training set $\mathcal{T}_{tr}$ contains N images with corresponding ground-truth label set $\mathcal{Y} = \{y_i\}_{i=1}^{N}$, where $y_i \in \{1, 2, \cdots, C\}$ and C refers to the number of classes in $\mathcal{T}_{tr}$. And the testing set contains query set and gallery set. Thus, $\mathcal{T}_{te} = \{\mathcal{T}_{query}, \mathcal{T}_{gallery}\}$.

Let $\mathbf{V}$, $\mathbf{I}$, and $\mathbf{X}$ denote the visible images, the infrared images, and the learned *X* modality images, respectively. We set f as the deep feature learner, and g as the lightweight *X* modality generator. Then we have

$$\mathbf{X} = g(\mathbf{V}). \tag{1}$$

During testing, we find the nearest visible neighbor $\mathbf{V}_{j^*} \in \mathcal{T}_{gallery}$ of a query image $\mathbf{I}_i \in \mathcal{T}_{query}$ from the gallery. And we generate $\mathbf{X}$ to assist cross-modal search. Thus, we get index j^* using:

$$j^* = \arg\min_j (D(f(\mathbf{I}_i), f(\mathbf{V}_j)) + D(f(\mathbf{I}_i), f(\mathbf{X}_j))), \tag{2}$$

where $D(\cdot)$ is the Euclidean distance.

The proposed X-Infrared-Visible cross-modal learning approach for IV-ReID is shown in Fig. 2. Through the *X* modality generation, the three infrared, *X*, and visible modalities are fed into a weight-sharing deep feature learner. With the extracted features and classification outputs, we adopt modality respective gap constraint for each one and cross modality gap constraint between infrared and visible modalities, as well as infrared and *X* modalities. In the following, we describe the *X* modality, weight-sharing feature learner, and modality gap constraints in detail.

X Modality

For visible image based deep models, the appearance and colour information usually dominate the learned high-level semantic information (Zhong et al. 2019; Sun and Zheng 2019). Conversely, infrared image has only one channel of invisible electromagnetic radiation, making the semantic structure and shape information major roles (Jungling and Arens 2010). The infrared and visible modalities inherently contain much different information. Previous approaches intended to learn the cross-modal information directly from these two original modalities. The evaluation performance shows that this kind of direct mapping is not good enough to narrow the gap between the two modalities (Wu et al. 2017; Dai et al. 2018; Wang et al. 2019c).

Comparatively, we learn an adjoint, auxiliary modality with self-supervision as an assistant to reconcile the infrared and visible modalities.

As shown in Fig. 2, the lightweight *X* modality generator uses visible images as input, siphons off knowledge from the visible and infrared modalities. This non-linear lightweight network contains two 1×1 convolutional layers and a ReLU (Krizhevsky, Sutskever, and Hinton 2012) activation layer. It first transforms the visible images to a one-channel images and then reconstructs three-channel images. The first 1×1 convolutional layer maps the original three-channel visible images to one-channel images like the infrared ones. A ReLU activation layer is provided to improve the non-linear representation capability. Then, another 1×1 convolutional layer is used to map the non-linear activated one-channel arrays to three-channel *X* modality images as the original visible ones.

The learned *X* images with automatically generated labels from visible images provide additional self-supervision information. Compared with other auxiliary structure used methods, like GAN-based ones (Dai et al. 2018; Wang et al. 2019c; 2019a), we implement a much more lightweight and efficient network. The network is easier to optimize than GANs as well. Additionally, the main modality gap between infrared and visible modalities remains in channel space. These GAN-based methods reconstructs information not only in channel dimension but also in the spatial dimension, destroying the original spatial structure information. Comparatively, we apply the more rational 1×1 convolutional layers to learn the *X* modality, which is only a reconstruction of channel-wise information from the visible modality.

Weight-sharing Feature Learner

With the designed *X* modality generator, we propose an X-Infrared-Visible cross-modal feature learner on the basis of an effective baseline model (Luo et al. 2019). As shown in Fig. 2, the feature learner takes three modalities, namely the *X*, infrared, and visible ones, as input, and learns the cross-modal information in a common feature space. By sharing weight for the three modalities, the proposed X-Infrared-Visible framework becomes much more compact.

The *X* modality plays a critical role in the framework, as it bridges the cross-modal information communication through a modality gap constraint, which will be introduced later. By jointly considering the infrared-X as well as infrared-visible cross-modal disparity, the *X* learns from visible modality and infrared modality. The *X* generator is lightweight and simplified, and could be directly integrated with the weight-sharing feature learner.

During training, we optimize these three modalities jointly. *X* modality acts as an adjoint and auxiliary assistant to ease the learning difficulty. During testing, we combine the similarities calculated with infrared-X and infrared-visible pairs as Eq. 2, and achieve the best performance.

Modality Constraints

Considering the infrared-visible cross-modal matching protocol during testing as introduced before, cross-modal constraints are the priority in optimizing (Ye et al. 2018b; Feng, Lai, and Xie 2019). Previous methods consider enhancing the feature discriminability with positive infrared-visible pairs and negative infrared-visible pairs. Comparatively, we form the modality gap constraints with infrared, visible, and *X* modalities in a joint manner. We form the well-aligned batches with the size of $3M$, of which the first M is for infrared modality, the second M for *X* modality, and the third M for visible modality. Thus, the cross modality gap (CMG) constraint $\mathcal{L}_C$ could be computed as follows:

$$\mathcal{L}_C = \mathcal{L}_{cross}^{\mathbf{I},\mathbf{X}} + \mathcal{L}_{cross}^{\mathbf{I},\mathbf{V}}. \tag{3}$$

Considering cross modality gap constraint between infrared modality and the learned *X* modality, the constraint is defined as follows:

$$\mathcal{L}_{cross}^{\mathbf{I},\mathbf{X}} = \frac{1}{M}(\mathcal{L}_{\mathbf{I}-\mathbf{X}} + \mathcal{L}_{\mathbf{X}-\mathbf{I}}), \tag{4}$$

where

$$\begin{aligned}\mathcal{L}_{\mathbf{I}-\mathbf{X}} = \sum_{i=1}^{M}[\alpha_1 + \max_{\substack{j=M+1,\cdots,2M \\ y_i=y_j}} D(f(\mathbf{I}_i), f(\mathbf{X}_j)) \\ - \min_{\substack{k=M+1,\cdots,2M \\ y_i\neq y_k}} D(f(\mathbf{I}_i), f(\mathbf{X}_k))]_+,\end{aligned} \tag{5}$$

and

$$\begin{aligned}\mathcal{L}_{\mathbf{X}-\mathbf{I}} = \sum_{i=M+1}^{2M}[\alpha_1 + \max_{\substack{j=1,\cdots,M \\ y_i=y_j}} D(f(\mathbf{X}_i), f(\mathbf{I}_j)) \\ - \min_{\substack{k=1,\cdots,M \\ y_i\neq y_k}} D(f(\mathbf{X}_i), f(\mathbf{I}_k))]_+.\end{aligned} \tag{6}$$

α_1 is a margin parameter and $[z]_+ = max(z, 0)$. With hard sample mining, the Euclidean distance of positive cross-modal pairs could be well optimized to be smaller than negative cross-modal pairs between infrared and *X* modalities. The same goes for cross modality gap constraint between infrared and visible modalities. Considering the evaluation protocol that infrared images are regarded as query images, the combined cross modality gap constraint targets at infrared images and forces the positive visible images and *X* images to approach infrared ones.

Additionally, for each modality, we also apply the modality respective gap constraint to help the model converge since the intra-modality constraints are easier than inter-modality constraints in convergence. Concretely, we apply cross entropy identity loss and an improved triplet loss to optimize the feature learning. The modality respective gap (MRG) $\mathcal{L}_M$ constraint is defined as follows:

$$\mathcal{L}_M = \mathcal{L}_{\mathbf{I}} + \mathcal{L}_{\mathbf{X}} + \mathcal{L}_{\mathbf{V}}. \tag{7}$$

Take infrared modality gap as an example, the combined loss constraint is defined as:

$$\mathcal{L}_{\mathbf{I}} = \frac{1}{M}(\mathcal{L}_{\mathbf{I}}^{id} + \mathcal{L}_{\mathbf{I}}^{tri}), \tag{8}$$

where

$$\mathcal{L}_{\mathbf{I}}^{id} = -\sum_{i=1}^{M} \mathbf{y}_i log(\mathbf{p}_i), \tag{9}$$

and

$$\mathcal{L}_{\mathbf{I}}^{tri} = \sum_{i=1}^{M}[\alpha_2 + \max_{\substack{j=1,\cdots,M \\ y_i=y_j}} D(f(\mathbf{I}_i), f(\mathbf{I}_j)) - \min_{\substack{k=1,\cdots,M \\ y_i \neq y_k}} D(f(\mathbf{I}_i), f(\mathbf{I}_k))]_+. \tag{10}$$

We set $\mathbf{p}_i$ here to represent the classification outputs of the image $\mathbf{I}_i$, and $\mathbf{y}_i$ to represent one-hot vector of label y_i. $\mathcal{L}_{\mathbf{X}}$ and $\mathcal{L}_{\mathbf{V}}$ can be defined analogously to $\mathcal{L}_{\mathbf{I}}$. We set margin α_2 to differentiate from the one used in $\mathcal{L}_C$ since the difficulties vary between inter-modality and intra-modality.

Optimization

The optimization of the proposed X-Infrared-Visible cross-modal learning framework integrated with an *X* modality could be directly conducted in an end-to-end manner, by cascading $\mathcal{L}_M$ and $\mathcal{L}_C$ as follows:

$$\mathcal{L} = \mathcal{L}_M + \lambda \mathcal{L}_C. \tag{11}$$

Here, λ is a trade-off hyperparameter for balancing the contributions of the two modality gap constraints.

Experiments

In this section, we compare the performance of the proposed method with other state-of-the-art methods and evaluate the contribution of the key components.

Experimental Settings

Datasets. We perform experiments on two publicly available datasets SYSU-MM01 and RegDB.

- **SYSU-MM01** is a challenging, large-scale dataset dedicated to infrared-visible cross-modal person ReID, collected by four visible cameras and two near-infrared ones (Wu et al. 2017). It contains in total 30,071 visible images and 15,792 infrared images of 491 identities, in which each identity is captured by one visible camera and one near-infrared camera at least. We use the *single-shot all-search mode* evaluation protocol (Wu et al. 2017), since it is the most challenging mode and adopted in all the comparative methods. The dataset is divided into a training set with 395 identities and a testing set with 96 identities. The training set consists of 22,258 visible images and 11,909 infrared images, while the query set and the gallery set contain 3,803 infrared images and 301 randomly sampled visible images, respectively.
- **RegDB** is constructed by using a pair of aligned visible and infrared cameras (Nguyen et al. 2017). It contains 412 identities with 10 visible images and 10 far-infrared images for each one. Following (Ye et al. 2018a; Wang et al. 2019c), we randomly split the dataset into two halves, one for training and the other for testing. Training set consists of 2,060 visible images and 2,060 infrared images. The same goes for testing set. Query set consists of 2,060 infrared images and gallery set contains 2,060 visible images.

Evaluation metrics. The Cumulative Matching Characteristics (CMC) curve and the mean Average Precision (mAP), as two standard evaluation metrics widely used in infrared-visible cross-modal person ReID works, are adopted in our experiments. For statistically stable results, the evaluation procedure is repeated for 10 trials with randomly sampled query and gallery sets, as (Ye et al. 2018a; Feng, Lai, and Xie 2019; Wang et al. 2019c).

Implementation details. The proposed method is implemented with PyTorch. We adopt the Adam optimizing method and the initial learning rate is set to be 0.00035 with warm-up strategy. The weight decay is set to be 0.0005. The batch size M for each modality is set to be 48 on one single TITAN Xp GPU, resulting a total mini-batch of 144. We set the training epoch to 120. And the learning rate decays at 40th and 70th epoch with a decay factor of 0.1. The margin parameter α_1 in $\mathcal{L}_C$ is set to be 0.5 while the margin parameter α_2 in $\mathcal{L}_M$ is set to be 0.3. As this task focuses more on cross-modal search, we set larger α_1 to emphasis cross modality constraints. The trade-off hyperparameter λ between two modality constraints in Eq. 11 is set to be 0.1.

For a comprehensive comparison, we choose two backbone networks, ResNet-50 (He et al. 2016) and DenseNet-121 (Huang et al. 2017), both pre-trained on ImageNet (Deng et al. 2009), to validate our method. ResNet-50 is a modern choice to provide state-of-the-art performance in person ReID (Wang et al. 2019a; 2019c; Luo et al. 2019). We adopt a modified ResNet-50 architecture (Luo et al. 2019). All training and testing images are re-scaled to a fixed size of 256×128 for ResNet-50. To better demonstrate the effectiveness of our each innovative component, we also perform the ablation study using DenseNet-121, where all images are re-scaled to a fixed size of 224×224.

Comparison with State-of-the-art Methods

We evaluate our method with other state-of-the-art methods. The first two listed in Table 1 are introduced at length in (Wu et al. 2017), which are the modifications of the widely used IDE method (Zheng, Yang, and Hauptmann 2016) under IV-ReID protocol. The remaining methods are dedicated to IV-ReID experimental settings as introduced before, including Zero-Padding (Wu et al. 2017), TONE (Ye et al. 2018a), HCML (Ye et al. 2018a), BDTR (Ye et al. 2018b), D-HSME (Hao et al. 2019), cmGAN (Dai et al. 2018), D^{2}RL (Wang et al. 2019c), MSR (Feng, Lai, and Xie 2019) and AlignGAN (Wang et al. 2019a).

Table 1 lists the comparison results on the SYSU-MM01 and RegDB datasets. The results demonstrate that the proposed approach for cross-modal person ReID outperforms existing state-of-the-art methods by a great margin on both two datasets. Compared with the state-of-the-arts, our method achieves an absolute gain of 7.5% and 10% at least in terms of rank 1 and mAP metric on SYSU-MM01 datasets. And our method outperforms other methods at least 5% and 7% in terms of rank 1 and mAP on RegDB dataset. Specifically, we achieve 62.21% rank 1 and 60.18% mAP accuracy on the RegDB dataset, and 49.92% rank 1 and 50.73% mAP accuracy on the SYSU-MM01 dataset.

Table 1: Comparison results (%) with the state-of-the-art IV-ReID methods on RegDB and SYSU-MM01 datasets.

Approach	RegDB				SYSU-MM01			
	$r = 1$	$r = 10$	$r = 20$	mAP	$r = 1$	$r = 10$	$r = 20$	mAP
One-stream (Wu et al. 2017)	13.11	32.98	42.51	14.02	12.04	49.68	66.74	13.67
Two-stream (Wu et al. 2017)	12.43	30.36	40.96	13.42	11.65	47.99	65.50	12.85
Zero-Padding (Wu et al. 2017)	17.75	34.21	44.35	18.90	14.80	54.12	71.33	15.95
TONE (Ye et al. 2018a)	16.87	34.03	44.10	14.92	12.52	50.72	68.60	14.42
HCML (Ye et al. 2018a)	24.44	47.53	56.78	20.80	14.32	53.16	69.17	16.16
BDTR (Ye et al. 2018b)	33.47	58.42	67.52	31.83	17.01	55.43	71.96	19.66
D-HSME (Hao et al. 2019)	50.85	73.36	81.66	47.00	20.68	62.74	77.95	23.12
cmGAN (Dai et al. 2018)	-	-	-	-	26.97	67.51	80.56	27.80
D^{2}RL (Wang et al. 2019c)	43.40	66.10	76.30	44.10	28.90	70.60	82.40	29.20
MSR (Feng, Lai, and Xie 2019)	48.43	70.32	79.95	48.67	37.35	83.40	93.34	38.11
AlignGAN (Wang et al. 2019a)	57.90	-	-	53.60	42.40	85.00	93.70	40.70
Our method	**62.21**	**83.13**	**91.72**	**60.18**	**49.92**	**89.79**	**95.96**	**50.73**

Table 2: Comparison with cmGAN, D^{2}RL using same backbone on the SYSU-MM01 dataset.

Method	$r = 1$	$r = 10$	$r = 20$	mAP
cmGAN	26.97	67.51	80.56	27.80
D^{2}RL	28.90	70.60	82.40	29.20
Ours	33.28	77.05	88.27	33.76

Table 3: Comparison with AlignGAN using same backbone on the SYSU-MM01 dataset.

Method	$r = 1$	$r = 10$	$r = 20$	mAP
AlignGAN	42.40	85.00	93.70	40.70
Ours	44.12	83.57	92.04	44.22

Compared with the relevant GAN-based methods cmGAN, D^{2}RL, and AlignGAN, the proposed *X* modality generator is not only more lightweight, also much more effective. Firstly, the generator in our method just has two 1×1 convolution layers, much less than the deep network 'UNIT' (over 20 layers) (Liu, Breuel, and Kautz 2017) in D^{2}RL. We also measure the processing FLOPs of the generator. For a 256×128 input image, the FLOPs of the generator are 0.0002G, which is trivial compared with the following feature extractor (2.7G). As shown in Table 2 and Table 3, our method improves a lot with same backbone as cmGAN, D^{2}RL, and AlignGAN used respectively. With modified high baseline network in (Luo et al. 2019), our method continuously improves the mAP and rank 1 accuracy to 50.73% and 49.92% on the SYSU-MM01 dataset.

Ablation Study

In this subsection, we evaluate the proposed components introduced before in detail on the large-scale SYSU-MM01 dataset. Some intermediate results are provided as follows to illustrate how much each of them contributes to the final significant performance. We use some acronyms for better illustration. Specifically, X is short for *X* modality and CMG is short for cross modality gap constraints.

- Baseline: Baseline model is trained using $\mathcal{L}_{\mathbf{V}} + \mathcal{L}_{\mathbf{I}}$.

Table 4: Ablation study on the SYSU-MM01 dataset.

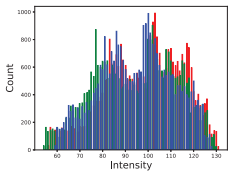
ResNet-50				
Method	$r = 1$	$r = 10$	$r = 20$	mAP
Baseline	38.39	81.65	90.84	40.62
Baseline+X	45.57	86.45	94.27	47.03
Baseline+X+CMG	49.92	89.79	95.96	50.73
DenseNet-121				
Method	$r = 1$	$r = 10$	$r = 20$	mAP
Baseline	38.67	81.66	90.18	39.68
Baseline+X	41.29	82.41	91.08	41.09
Baseline+X+CMG	48.20	88.57	94.84	48.01

- Baseline+X: This model is integrated with the proposed *X* modality, and is trained using $\mathcal{L}_{\mathbf{V}} + \mathcal{L}_{\mathbf{I}} + \mathcal{L}_{\mathbf{X}}$.
- Baseline+X+CMG: The proposed whole framework is optimized with final integrated loss function $\mathcal{L}$ as Eq. 11.

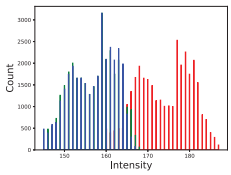
We adopt two commonly used backbone model ResNet-50 and DenseNet-121, to evaluate each component we provided. As shown in Table 4, the effectiveness of each component is clearly revealed. The learned *X* modality improves the baseline model with 3-7% rank 1 and 2-7% mAP accuracy on the SYSU-MM01 dataset. And the cross modality gap constraint continuously improves the model with *X* 4-7% rank 1 accuracy and 3-7% mAP performance. Our self-supervised *X* modality has significantly eased the learning difficulties in cross-modal matching just with a lightweight network and slight costs.

Discussions

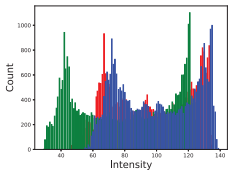
A closer look at *X*. We statistically analyze the average single-color-channel intensity of all pixels inside an image over the training images of the SYSU-MM01 and RegDB datasets. Fig. 3 shows the histograms computed from the natural visible images and the *X* modality images of the two datasets, respectively. The statistics of the three color channels of natural visible images are analogous. However, the 'R' channel in the *X* images has much higher intensity than 'G' and 'B'. As shown in Fig. 4, compared with natural visible images, the *X* images appear much 'redder' and



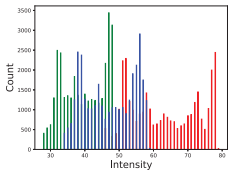
(a) "Visible" of SYSU-MM01



(b) "*X*" of SYSU-MM01



(c) "Visible" of RegDB



(d) "*X*" of RegDB

Figure 3: Histograms of average single-colour-channel intensity of all pixels inside an image over the training sets of SYSU-MM01 and RegDB. They are computed from visible and *X* images respectively. Red, green, and blue histograms represent the corresponding 'R', 'G', and 'B' channels. (**Best viewed in color.**)

Figure 4: Visualization of the three modalities. (Best viewed in color.)

thus have longer wavelength from the viewpoint of electromagnetic radiation. Conversely, when compared with infrared images, the *X* images look more colorful and thus the wavelength shorten. As a result, the proposed method finally learns a 'new' modality with an intermediate wavelength between visible and infrared light. With the *X*, cross-modal learning becomes much easier as it is avoided to deal with the gap between the visible and the infrared modalities independently. As shown in Fig. 1, cross-modal matching can be more easily performed with the feature learned with *X* modality than normal IV-ReID baseline model.

It is interesting to evaluate the performance of several special cases of the *X* modality generator g. The comparison results are reported in Table 5, where 'Mean' refers to channel-wise average of visible images, 'Gray' refers to gray information extracted from visible images, 'V' refers to the lightness channel V information of visible images in HSV colour space and 'Y' refers to the Y channel information of visible images in YCbCr colour space. The extracted one-channel information is expanded three times in channel dimension and fed into the weight-sharing feature learner. As discussed in digital camera sensors works (Fredembach and Süsstrunk 2008), in creating 'pleasing' images, the 'V' and 'Y' channel information show closer performance as near infrared information. Similarly, the comparison results in Table 5 show that 'V' and 'Y' channel information are better than 'Mean' and 'Gray' information in helping to match infrared images with visible ones. The proposed lightweight generator clearly outperforms all special cases.

Table 5: Performance of special cases of '*X*' using ResNet-50 model on the SYSU-MM01 dataset.

Method	$r = 1$	$r = 10$	$r = 20$	mAP
Baseline+Mean	41.73	83.89	92.04	44.27
Baseline+Gray	40.34	82.66	91.07	42.92
Baseline+V	44.61	84.40	93.51	46.49
Baseline+Y	43.69	84.52	93.28	45.86
Baseline+X	45.57	86.45	94.27	47.03

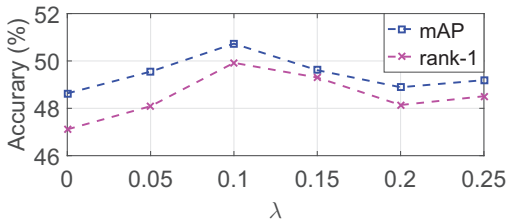


Figure 5: Performance of rank 1 and mAP accuracy with respect to the hyperparameter λ on the SYSU-MM01 dataset.

Parameter influence. We evaluate the influence of the trade-off hyperparameter λ between modality respective gap constraint $\mathcal{L}_M$ and cross modality gap constraint $\mathcal{L}_C$ used in Eq. 11. The $\mathcal{L}_M$ is easier than $\mathcal{L}_C$ in terms of convergence. While the target of the IV-ReID task is to learn cross-modal matching and narrow the cross modality gap. Fig. 5 shows the performance of rank 1 and mAP based on modified ResNet-50 with SYSU-MM01 dataset by varying the parameter λ. We could find that cross-modal person ReID performance varies when the trade-off hyperparameter changes. And there exists a suitable value to balance the $\mathcal{L}_M$ and $\mathcal{L}_C$. $\mathcal{L}_M$ helps convergence of the model and $\mathcal{L}_C$ helps the cross-modal information learning.

Conclusion

This paper focuses on the infrared-visible cross-modal person re-identification task. To mitigate the inherent modality gap between the infrared and visible modalities, we propose a new X-Infrared-Visible (XIV) cross-modal learning framework with an adjoint and auxiliary *X* modality. Concretely, we design a lightweight generator to siphon off knowledge from the visible and infrared modalities, and output the *X* modality images. Then, a weight-sharing deep feature learner is provided to extract cross-modal features and classification outputs in a joint manner. We optimize the generator and feature learner directly with the devised modality respective gap (MRG) constraint and cross modality gap (CMG) constraint in an end-to-end manner. Experimental results on two publicly available infrared-visible cross-modal person re-identification datasets SYSU-MM01 and RegDB demonstrate the superiority of the proposed *three-*

mode cross-modal learning approach.

Acknowledgments

This work was supported by National Basic Research Program of China (Grant No.2015CB351705) and National Major Project (Grant No.2017YFC0803905).

References

Agrawal, P.; Carreira, J.; and Malik, J. 2015. Learning to see by moving. In *ICCV*.

Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; and Zheng, N. 2016. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. IEEE.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *ICCV*.

Feng, Z.; Lai, J.; and Xie, X. 2019. Learning modality-specific representations for visible-infrared person re-identification. *TIP* 29:579–590.

Fredembach, C., and Süsstrunk, S. 2008. Colouring the near-infrared. In *CIC*, volume 2008, 176–182. Society for Imaging Science and Technology.

Gomez, L.; Patel, Y.; Rusinol, M.; Karatzas, D.; and Jawahar, C. V. 2017. Self-supervised learning of visual features through embedding images into text topic spaces. In *CVPR*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*, 2672–2680.

Hao, Y.; Wang, N.; Li, J.; and Gao, X. 2019. Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In *AAAI*, volume 33, 8385–8392.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*, 4700–4708.

Jungling, K., and Arens, M. 2010. Local feature based person reidentification in infrared image sequences. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 448–455. IEEE.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *NIPS*, 700–708.

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 *CVPR Workshops*, 0–0.

Ma, Z.; Wei, X.; Hong, X.; and Gong, Y. 2019. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, 6142–6151.

Nguyen, D.; Hong, H.; Kim, K.; and Park, K. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors* 17(3):605.

Patel, Y.; Gomez, L.; Rusiñol, M.; Karatzas, D.; and Jawahar, C. 2019. Self-supervised visual representations for cross-modal retrieval. In *ICMR*, 182–186. ACM.

Sun, X., and Zheng, L. 2019. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*.

Sun, Y.; Xu, Q.; Li, Y.; Zhang, C.; Li, Y.; Wang, S.; and Sun, J. 2019. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *CVPR*.

Tay, C.-P.; Roy, S.; and Yap, K.-H. 2019. Aanet: Attribute attention network for person re-identifications. In *CVPR*.

Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; and Gong, Y. 2010. Locality-constrained linear coding for image classification. In *CVPR*, 3360–3367. Citeseer.

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*.

Wang, X.; Huang, Q.; Celikyilmaz, A.; Gao, J.; Shen, D.; Wang, Y.-F.; Wang, W. Y.; and Zhang, L. 2019b. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*.

Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.-Y.; and Satoh, S. 2019c. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. Rgb-infrared cross-modality person re-identification. In *ICCV*, 5380–5389.

Yang, Q.; Yu, H.-X.; Wu, A.; and Zheng, W.-S. 2019. Patch-based discriminative feature learning for unsupervised person re-identification. In *CVPR*.

Ye, M.; Lan, X.; Li, J.; and Yuen, P. C. 2018a. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*.

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.

Zhang, S.; Wang, J.; Wang, Z.; Gong, Y.; and Liu, Y. 2015. Multi-target tracking by learning local-to-global trajectory models. *Pattern Recognition* 48(2):580–590.

Zheng, L.; Yang, Y.; and Hauptmann, A. G. 2016. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; and Yang, Y. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 598–607.