International Journal of Automation and Computing
www.ijac.net

18(3), June 2021, 443-456
DOI: 10.1007/s11633-020-1262-z

Learning Deep RGBT Representations for Robust Person Re-identification

Ai-Hua Zheng Zi-Han Chen Cheng-Long Li Jin Tang Bin Luo

Anhui Provincial Key Laboratory of Multi-modal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China

Abstract: Person re-identification (Re-ID) is the scientific task of finding specific person images of a person in a non-overlapping camera networks, and has achieved many breakthroughs recently. However, it remains very challenging in adverse environmental conditions, especially in dark areas or at nighttime due to the imaging limitations of a single visible light source. To handle this problem, we propose a novel deep red green blue (RGB)-thermal (RGBT) representation learning framework for a single modality RGB person Re-ID. Due to the lack of thermal data in prevalent RGB Re-ID datasets, we propose to use the generative adversarial network to translate labeled RGB images of person to thermal infrared ones, trained on existing RGBT datasets. The labeled RGB images and the synthetic thermal images make up a labeled RGBT training set, and we propose a cross-modal attention network to learn effective RGBT representations for person Re-ID in day and night by leveraging the complementary advantages of RGB and thermal modalities. Extensive experiments on Market1501, CUHK03 and DukeMTMC-reID datasets demonstrate the effectiveness of our method, which achieves state-of-the-art performance on all above person Re-ID datasets.

Keywords: Person re-identification (Re-ID), thermal infrared, generative networks, attention, deep learning.

Citation: A. H. Zheng, Z. H. Chen, C. L. Li, J. Tang, B. Luo. Learning deep RGBT representations for robust person re-identification. *International Journal of Automation and Computing*, vol.18, no.3, pp.443–456, 2021. http://doi.org/10.1007/s11633-020-1262-z

1 Introduction

Person re-identification (Re-ID) aims to match pedestrian images with the same identity across multiple non-overlapping cameras. It has been a hot research topic since the last decade with potential practical applications like video surveillance and pedestrian retrieval for public security. However, person Re-ID encounters many challenges, such as the cross-view changes of a person′s pose, illumination, viewpoints, backgrounds clutter, occlusion, etc.

Extensive efforts have been made in the past decade to overcome these challenges. Early works dedicated to either feature extraction[1–3] or metric learning[3,4] schemes. Feature extraction based methods aim to learn the discriminative features to maintain invariance of the same person, and the distinction among different persons. Metric learning based approaches mainly train a distance measurement or a classifier to solve the intra-class discrepancy and inter-class similarity. With the extensive applications of deep learning[5], convolution neural networks (CNNs) have been widely used in person Re-ID to automatically learn more discriminative features[3, 6–16]. These methods mainly employ deep classification models to learn discriminative feature representations for the visual appearance of person images. Meanwhile, deep metric learning methods are also widely implemented in person Re-ID[17–24] by minimizing the inter-class diversities and maximizing the intra-class distinctions. Recently, with the blossom of generative adversarial networks (GANs), some researchers have tried to use the generation models to relieve the pose and camera style variations across the cameras in person Re-ID[25–28].

Other researchers have focused on the temporal information or the gait information[29, 30] for video based person Re-ID. Despite of the great progress on person Re-ID, most of the existing single red green blue (RGB)-modality Re-ID methods and benchmark datasets are based on favorable lighting, which limits their capability in real-life applications in the adverse environments, such as poor illumination in bad weather or at nighttime.

Thermal infrared (TIR) cameras can capture infrared radiation emitted by subjects with a temperature above absolute zero[31]. These cameras are insensitive to lighting conditions and have a strong ability to penetrate haze and smog. Therefore, the advantage of thermal images is that they are not affected by low illumination, illumination changes and shadows, as shown in Fig. 1.

Research Article
Manuscript received June 22, 2020; accepted October 10, 2020; published online January 19, 2021
Recommended by Associate Editor Jangmyung Lee
Colored figures are available in the online version at https://link.springer.com/journal/11633

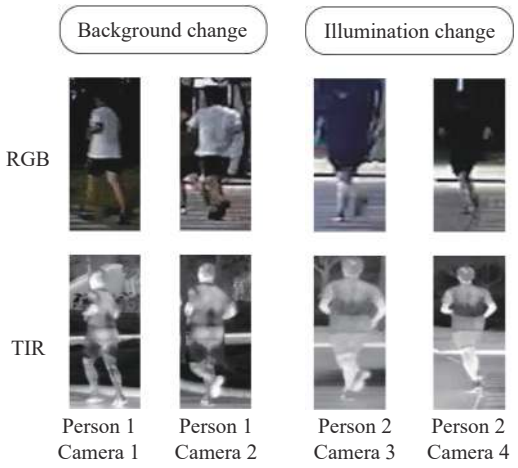


Fig. 1 Advantage of TIR images compared to RGB images. Due to the illumination change or background clutter, the same person may appear differently under the visible RGB cameras, while the TIR images attenuate the influence in these challenging scenarios.

Recently, as the quality of TIR images has improved and the cost of infrared cameras has reduced, TIR data has been widely used in computer vision tasks to overcome the limitations in conventional RGB environments, such as RGB-thermal (RGBT) tracking[31,32] and RGBT object detection[33], which take advantage of the characteristics of the two modalities to improve the performance of the corresponding tasks. Meanwhile, Nguyen et al.[34] proposed a RGBT dataset, which contains RGB and thermal infrared image pairs for pedestrian recognition. This dataset has been widely used for cross-modality person Re-ID[35–37].

Although TIR data can relieve the challenge in adverse conditions in a conventional RGB single modality, most of the surveillance environments are based on single visibility only, which results in the lack of TIR data resources for the RGBT Re-ID task. Therefore, how to utilize the advantage of the thermal infrared modality for the single RGB modality Re-ID is still an open problem.

In recent years, many researchers employ GANs in cross-modal generation to supplement the single modal information. For instance, Zhang et al.[38] used the image translation method to generate thermal images in thermal infrared tracking. Inspired by these works, by training on existing RGBT datasets, we propose to use the generative adversarial network to translate the labeled RGB person images to thermal infrared ones. The labeled RGB images and the synthetic thermal images consist of the labeled RGBT training set, which makes use of the complementary information from both visible and thermal modalities. Specifically, we employ CycleGAN[39] which was proposed to learn mappings between unpaired domains for the cross-modal generation.

After obtaining the synthetic RGBT training dataset, we can learn both RGB and thermal representations with the input of a single RGB-modality query during the test. The forthcoming issue is how to effectively fuse the information from both modalities. The intuitive approach is to concatenate the representations from each modality. However, different modalities may contribute unequally in different scenarios. Recently, attention models[40] have drawn much attention and been successfully applied to all kinds of visual mechanisms, such as pedestrian counting[41], action recognition[42], video summarization[43] and object detection[44]. In our task, we propose to employ a channel-spatial attention network[45] to learn more discriminative RGBT representations to further boost the performance.

Based on the above discussion, we propose a novel deep RGBT representation learning framework for single RGB person re-identification. First, we synthesize the thermal person images via CycleGAN[39] for RGB person images. Then, we learn the RGBT representation based on the synthetic RGBT data. Finally, we employ the attention network to improve the representation of the network and balance information between the two modalities. By exploiting multi-modal representation with the proposed method, it can relieve the illumination and background clutter in conventional RGB Re-ID tasks while making full use of the complementary information from both RGB and thermal modalities without additional modality resources. The contributions of this paper can be summarized as follows:

1) We propose a deep RGBT representation learning framework for single RGB-modality person Re-ID. By transferring the RGB query images to TIR ones, our method can take the advantages of both RGB and thermal modalities without additional modality resources in RGB person Re-ID.

2) We propose to employ the channel-spatial attentions in our network to automatically learn the important information when fusing RGB and thermal representations for robust RGB person Re-ID.

3) Extensive experiments on prevalent RGB person Re-ID datasets, including Market1501[18], DukeMTMC-reID[8] and CUHK03[6], show the promising performance of the proposed method especially in adverse scenarios.

2 Related work

2.1 RGB person re-identification

Person Re-ID has been attracting more attention in recent years. Early approaches focused on extracting hand-crafted features. Representative descriptors include histograms of oriented gradients (HOG)[1], local maximal occurrence (LOMO)[3] and local binary patterns (LBP)[2]. Meanwhile, metric learning based methods emerged for

learning the optimized subspace to minimize the cross-view gap in person Re-ID, such as keep it simple and straightforward metric (KISSME)[4], cross-view quadratic discriminant analysis (XQDA)[3] and top-push[20]. The development of CNNs accelerates the recent progress in person Re-ID. Chen et al.[46] proposed CNN structures to extract characteristic features with identification loss and verification loss. Zhao et al.[47] designed a deep CNN network named as spindle net to fuse global body features and body region features for person Re-ID. Su et al.[48] used a pose-driven deep CNN model to leverage the human part cues to alleviate the pose variations and learn robust features of global and local information. Sun et al.[49] proposed a part-based network part-based convolutional baseline (PCB) which learns more discriminative feature representations in the part level. Chang et al.[19] proposed a semantic level network that factorizes the visual appearance of a person. Ding et al.[13] proposed a feature mask network to re-weight different parts of high-level and low-level features.

Some researchers integrated the idea of metric learning into the deep CNNs for person Re-ID. Yi et al.[17] first combined the metric learning method (cosine distance) with deep CNN. Chen et al.[50] designed a quadruplet loss leading to large inter-class variation and smaller intra-class variation than triplet loss. Yao et al.[24] improved the performance by computing the person classification loss on each part separately. Zhu et al.[23] proposed a network to learn the distance metric by designing different objective functions for hard and easy negative samples. Yuan et al.[51] proposed a fast-approximated triplet (FAT) loss to preserve the effectiveness of triplet loss.

Recently, many researchers paid attention to GAN-based methods for person Re-ID. Zheng et al.[8] proposed to generate unlabeled samples with a simple semi-supervised pipeline on the original training dataset, and adopted the deep convolution generative adversarial network (DCGAN) for data generation.

Person transfer generative adversarial network (PT-GAN)[26] is proposed to address the problem of poses variation in person Re-ID where the model is trained with rich pose variations which are generated via transferring pose instances. Ge et al.[27] proposed feature distilling generative adversarial network (FD-GAN) to learn pose-unrelated person features with pose guidance. Wu et al.[14] used adversarial learning to address the view discrepancy by optimizing the cross-entropy view confusion objective in person Re-ID. However, person Re-ID on a single RGB modality faces big challenges with illumination changes especially for dark lighting conditions in the severe weather or night-time.

2.2 Multi-modal person re-identification

Recently, with the development of multi-modal vision, multi-modal person Re-ID has gained much attention. Barbosa et al.[52] proposed a pattern analysis and computer vision (PAVIS) dataset, which contains two groups of RGB and depth person images. Munaro et al.[53] proposed a BIWI dataset, which consists of 50 different persons in RGBD data. Nguyen et al.[34] proposed a RegDB dataset which contains 4 120 RGB and thermal person image pairs for person recognition.

Based on the above RGBD person Re-ID datasets, Pala et al.[54] combined clothing appearance with depth data for person Re-ID. Mogelmose et al.[55] proposed a tri-modal (RGB, depth, thermal) person Re-ID to combine RGB, depth and thermal features. Xu et al.[56] proposed a distance metric using RGB and depth data to improve RGB-based person Re-ID. John et al.[57] combined RGB-height histogram and gait features of depth information for person Re-ID.

Wu et al.[58] proposed a kernelized implicit feature transfer scheme to estimate the Eigen-depth feature from RGB images implicitly when the depth device was not available. Paolanti et al.[59] combined depth and RGB data with multiple k-nearest neighbor classifiers based on different distance functions. Ren et al.[60] exploited a uniform and variational deep learning method for RGBD object recognition and person Re-ID. However, most of existing surveillance systems are based on single RGB camera networks, and thus how to utilize the advantages of the thermal infrared modality in single RGB-modality person Re-ID is still an open question.

2.3 Cross-modal generation

Generative adversarial networks (GANs)[61] have achieved great success recently, especially in image generation[62, 63], image editing[64] and image-to-image translation[39, 65]. Conditional GANs (cGAN)[62] have been proposed based on a selected input variable. With the rise of cross-modal simulation research, in recent years cross-modal image generation has attracted much attention. Xu et al.[66] proposed a method to reconstruct thermal images from the associated RGB data and learn cross-modal deep representations for detection. Zhang et al.[38] used the generative adversarial network of style transfer to generate thermal infrared images from visible images to alleviate the thermal tracking problem in weak illumination. Cross-modal generation has also performed well in other areas. Luo et al.[67] combined binocular images with monocular images to generate depth modality images of monocular images. Qiao et al.[68] proposed a novel global-local model to generate images from texts. Chen et al.[69] exploited conditional GANs to achieve the generation of audio-images. Zhou et al.[70] enabled arbitrary-subject talking face generation by learning disentangled audio-visual representations. The above progress on cross-modal generation provides another way to make use of the advantages of the thermal infrared modality in person Re-ID with a single RGB modality.

3 Proposed approach

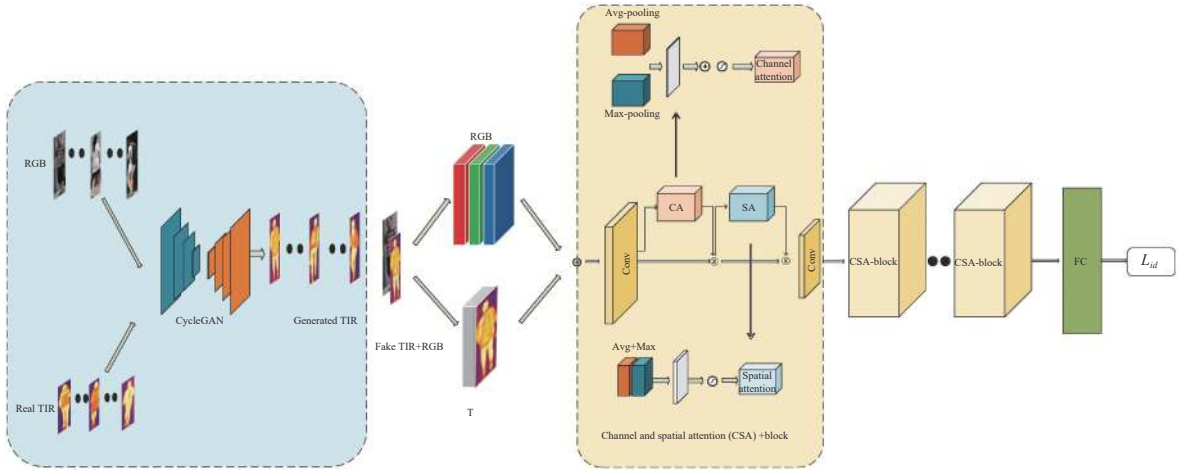
3.1 Overview of our approach

Our proposed framework is shown in Fig. 2. We aim to leverage the thermal information to boost the traditional RGB Re-ID in challenging scenarios. There are two main parts in our approach, including 1) thermal data generation networks, which transfer the RGB images into TIR ones via CycleGAN[39], and 2) attentive RGBT Re-ID network, which utilize the channel attention (CA) and the spatial attention (SA)[45] to highlight the meaningful information of input RGB-TIR image pairs for the Re-ID task. We shall elaborate the details of each module in the following two subsections.

3.2 Thermal data generation network

3.2.1 Network architecture

Image-to-image translation has been extensively researched in recent years. Representative translation methods include pix2pix[65], CycleGAN[39], etc. As we known, pix2pix[65] requires the paired input data. Due to



(a) Thermal data generation network (b) Attentive RGBT Re-ID network

Fig. 2 Framework of the proposed approach: (a) Thermal data generation network[39], generating a large TIR person image dataset. Both TIR and RGB data are used as input while training the generation model. After translating RGB data into TIR data, we acquire the RGB data together with the generated TIR data for future Re-ID task. (b) Attentive RGBT Re-ID network, learning RGBT representation based on both RGB and generated TIR data for Re-ID task. Colored figures are available in the online version.

Fig. 3 Samples of generated TIR images via CycleGAN and pix2pix for the corresponding RGB ones from Market1501[18]

the low quality RGB images in paired RGBT datasets RegDB[34], pix2pix[65] tends to generate blurring low quality data as shown in Fig. 3. Therefore, we employ the more advanced unpaired generation network CycleGAN[39] for better RGB to TIR transformation.

CycleGAN[39] is an effective method in image translation between two domains when the paired images are not available. Based on generative adversarial networks (GANs)[61], CycleGAN[39] consists of two generators and two discriminators, which mutually map an image from a source domain to a target domain with the cycle consistency loss. The generator in CycleGAN[39] contains two convolutions, six residual blocks and two fractionally-strided convolutions as the generator. The image generator G takes the encoded RGB person image features as inputs, and aims to decode new TIR person images. The discriminator is a convolutional PatchGAN[71], which distinguishes the decoded TIR image patches as real or fake. Let x be an image from source RGB domain X and y be an image from the target thermal infrared domain Y. Our target is to learn the mapping functions between RGB domain X and thermal domain Y. First, the adversarial loss is defined as the objective function,

$$\mathcal{L}_{GAN}(G, D_Y, X, Y) = E_{y\sim p(y)}\left[\log D_Y(y)\right] + E_{x\sim p(x)}[\log(1 - D_Y(G(x)))] \quad (1)$$

where the generator G is to generate images $G(x)$ that could transfer the style from source RGB domain X to thermal domain Y. The discriminator D_Y tries to distinguish whether the generated thermal images $G(x)$ are real or fake ones.

Additionally, the main idea of CycleGAN[39] is to introduce a cycle consistency loss, which maps the target domain Y back to source domain X. Therefore, unlike the conventional generative adversarial networks which only contain one generator, CycleGAN[39] includes another generator F to map $Y \rightarrow X$. The cycle consistency loss is defined as

$$\mathcal{L}_{cyc}(G, F) = E_x\left[||F(G(x)) - x||_1\right] + E_y\left[||F(F(y)) - y||_1\right]. \quad (2)$$

The cycle consistency loss makes the reconstructed images closer to the input images. In the same manner as the minimizing-and-maximizing game in traditional adversarial learning, the final objective function of CycleGAN[39] is defined as

$$G^*, F^* = \arg\min_G \max_D \left[\mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda\mathcal{L}_{cyc}(G, F)\right]. \quad (3)$$

Fig. 3 demonstrates several generated thermal samples in RGB dataset Market1501[18]. Some person images in the RGB modality are disturbed by background and illumination, especially as Person 2, Person 4 and Person 6 highlighted in red boxes. For instance, the third image of Person 2 captured outdoors is significantly disturbed by the background clutter compared with the first two indoor images. Similar background changes are also ubiquitous for other person images such as Person 4 and Person 6. While the corresponding generated TIR images can overcome these issues, compared with the pix2pix generation method, CycleGAN achieves more realistic synthesizing with much better visualized and more detailed appearance information.

3.2.2 Implementation details

We shall elaborate the data and training details to transfer the RGB person images into TIR ones in this section.

Data preparation. To transfer the RGB person images into TIR ones, the first requirement is the training data with both RGB and TIR person images. Currently, there are only two RGB-IR person Re-ID datasets, SYSU-MM01[72] and RegDB[34]. SYSU-MM01[72] is the prevalent RGB-NIR (near infrared) cross-modal dataset which captures with six individual non-overlapping cameras including four RGB ones and two NIR ones. However, NIR data is sensitive to the illumination and contains less information than TIR data, thus this dataset is not suitable for our work. RegDB[34] contains a large number of RGB-TIR image pairs which are captured by a binocular RGB-TIR camera set. Therefore, we train our cross-modal generation model on RegDB[34] with its TIR data in our paper.

As shown in Table 1, RegDB[34] contains 4 120 RGB-TIR image pairs of 412 identities under different lighting conditions. Each identity contains 10 different RGB-TIR image pairs. To make the generation of our method better, we choose high-quality person TIR images in different environments which have different poses in RegDB dataset, amounting to 2 154 TIR high-quality images for training. In order to relieve the influence of the unclear boundary in high temperatures, we select some data containing the challenge of unclear boundary conditions to train our generation network. The purpose of our approach is to integrate the generated thermal infrared information to the existing Re-ID datasets, such as Marker-1 501[18], CUHK03[6], and DukeMTMC-reID[8] datasets. We select 2 154 RGB person images from these datasets for training, which contains complex background and different poses. The generator is to learn more details about

Table 1 Datasets used for training the generation model and Re-ID task. We test our models at three single-modal datasets.

Types	Datasets	Number of images	
		RGB	IR
Multi-modal dataset	RegDB	4 120	4 120
Single-modal dataset	Market1501	32 217	–
	CUHK03	13 164	–
	DukeMTMC-reID	36 411	–

single-modal RGB data. For testing, we translate three RGB datasets to thermal infrared style, these amounts to total of about 10K images.

Training details. We train our generation network from scratch, initializing the weight from a Gaussian distribution with zero mean and standard deviation of 0.02. We empirically set $\lambda = 10$ in (3), and use the Adam method[73] to optimize the model with a batch size of 1. We use the same network architecture as in CycleGAN[39]. The input images have been resized as 128×128 pixels. We train CycleGAN[39] for 30 epochs with a learning rate of 0.000 2. In the first 20 epochs, we keep the same learning rate and decay the rate to zero in the next 10 epochs.

3.3 Attentive RGBT Re-ID network

3.3.1 Cross-modal convolution module

After obtaining the TIR data, the next step is to fuse the multi-modality information[74]. To leverage the TIR information to complement the conventional RGB Re-ID task, this section elaborates our cross-modal convolution module which aims to learn both RGB and TIR person features. We first encode each modality to a feature map of size $H \times W \times C$, where W and H indicate the feature dimensions, and C denotes the number of channels. As shown in Fig. 2, unlike the common convolution operation taking 3-channel RGB data as inputs, we input 4-channel RGBT data, including three channels RGB data I_{RGB} and one channel thermal data I_T:

$$I = I_{RGB} + \alpha I_T \tag{4}$$

where α is the balance parameter indicating the weight/contribution of the generated thermal data, and we empirically set this to 1 in this paper. "+" is operation of concatenation. We utilize the ResNet-50[75] as our backbone. We keep the layers of ResNet-50[75] till the Pooling-5 layer as the base network and change the dimension of the fully connected layer to N, which indicates the number of identities in the training dataset. We add a new embedding layer followed by linear and batch normalization[76], and then randomly crop it into a 256×128 rectangular image, each of which is flipped horizontally with 0.5 probability. The model has an additional data loader for TIR data to obtain the generated TIR images. It outputs the ID prediction logits p to calculate the cross-entropy loss.

Since the label of the generated TIR images are known, we calculate the difference between ID prediction logits p and the real labels. The cross-entropy loss is used to optimize the network, and formulated as

$$\mathcal{L}_{id}(a, n) = -\frac{1}{n}\sum_{i=1}^{n} \log p(b_i|a_i) \tag{5}$$

where n is the number of synthetic images in a training batch. p is the predicted probability of the input image a_i belonging to identity b_i. In general, the contributions of the generated thermal images and RGB images are different in different scenarios. To learn the diverse contribution of each modality, we further propose to employ the channel and spatial attention mechanisms to emphasize the discrimination.

3.3.2 Channel attention module

The convolutional block attention model[45] aims to produce a weighting map to carry out attention computation across the feature maps. In this way, the channel attention module will be oriented toward more important channels of the RGB-T feature.

Given an input RGB-T person image, we first obtain the feature map M from the first convolution layer of ResNet-50. The framework of the attention module is shown in Fig. 2. In particular, the network which embeds the channel attention module can be denoted as

$$M' = A_c(M) \otimes M \tag{6}$$

where $\otimes$ denotes the weighting operator. A_c is the channel attention module. The attention map A is expected to focus higher on a person region contrary to the background. M' indicates the channel attention features.

The channel attention module exploits the channel relationship of features by choosing the more meaningful channels of an RGB-T feature map. One can achieve the attention map via aggregating the input feature maps. A common way of aggregation is to use average-pooling to learn the extent of the input object. To better select the discriminative feature and preserve more texture information, we further introduce a max-pooling operation in this paper.

Both average-pooling and max-pooling[45] descriptors are forwarded to a convolution block to achieve the channel attention map. Finally, we use the element-wise summation to merge the output features. The objective functions can be defined as

$$\begin{aligned} A_c(M) =& M_{\text{Avg}} + M_{\text{Max}} = \sigma\left(C_2(\,ReLU\,(C_1\text{Avg}(\,M\,))\right)+ \\ & C_2(ReLU\,(C_1\text{Max}(\,M\,))\,)\,) \end{aligned} \tag{7}$$

where M denotes the feature of the image after different layers, σ denotes the sigmoid functions, C_1 and C_2 are two different convolution layers.

3.3.3 Spatial attention module

To capture the spatial relationship of features, we further employ the spatial attention module to emphasize the informative part of the features, as the complementary information to the channel attention. The spatial attention module in the network can be denoted as

$$M'' = A_s\left(M'\right) \otimes M' \tag{8}$$

where M' and M'' indicate features after the channel attention module and final attention features respectively. A_s is the spatial attention module.

Unlike channel attention, we concatenate two descriptors to generate an efficient descriptor and then forward to a convolution layer to compute the spatial attention map. The spatial attention is computed as

$$A_s(M) = M_{\text{Avg,Max}} = \sigma\left(C^{7\times7}\left(\text{Avg}(M) + \text{Max}(M)\right)\right) \quad (9)$$

where σ denotes the sigmoid function, and M indicates the feature of the image after different layers. $C^{7\times7}$ indicates the kernel size of the convolution layer. Fig. 4 demonstrates several samples enhanced by the channel-spatial attention map.

3.3.4 Implementation details

The backbone of our attentive RGBT Re-ID network is the standard ResNet-50[75]. The Re-ID network is trained with cross-entropy loss. The learning rate is set to 0.1 and then reduced to 0.01 at the 60th epoch. For the attention module, we use convolution layer with a kernel size of 7. The kernel size of spatial attention is 7×7.

We train our model in 90 epochs with an adjustable learning rate which will decrease while the epochs increase. We set the dropout to 0.5 to prevent overfitting. We use stochastic gradient descent (SGD) with the momentum of 0.9 and weight decay of 0.000 5 to fine-tune the network. All input images are resized to 256 × 128 with horizontal flipping during training.

4 Experimental results

To verify the effectiveness of our proposed method, we evaluate the method on three large-scale person Re-ID benchmark datasets, including Market1501[18], DukeMTMC-reID[8] and CUHK03[6]. Performance is evaluated by the cumulative matching characteristic (CMC) and mean average precision (mAP).

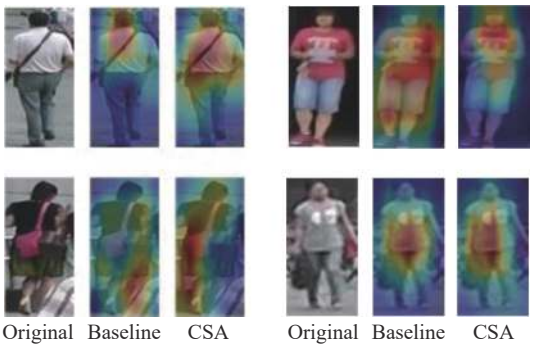


Fig. 4 Visualized feature maps of corresponding person images from the Market1501 dataset[18]. The results of baseline and CSA are achieved via ResNet-50 and ResNet-50 with the channel-spatial attention respectively on the original RGB data.

4.1 Datasets

Market1501[18] consists of 12 936 images of 751 identities for training and 19 281 images of 750 identities for testing from 6 camera views. There are on average 17.2 images per identity in the training set. In testing, 3 368 images from 750 identities are used as queries to retrieve the matching persons in the dataset.

DukeMTMC-reID[8] is a subset of tracking dataset DukeMTMC[77] for image-based person Re-ID. The dataset contains 16 522 images of 702 identities for training collected from 8 cameras and 2 228 query images from the other 702 identities. The evaluation metrics of the dataset is the same as that of Market1501[18].

CUHK03[6] contains 14 097 training images of 1 467 identities captured from two cameras where the scenario is the Chinese University of Hong Kong (CUHK) campus. Image samples of CUHK03 from 767 identities are selected for training, and the remaining 700 identities for testing.

4.2 Comparison with state-of-the-art methods

We first compare the performance of our method with the recent state-of-the-art Re-ID methods including some recent methods on three benchmark datasets.

4.2.1 Comparison on Market1501 dataset

Table 2 reports the comparison results on Market1501 datasets[18]. As we can see, our method beats the state-of-

Table 2 Experimental comparison of the proposed approach with state-of-the-art methods on Market1501[18] (in %)

Methods	Market1501		
	Rank-1	mAP	References
Bow+KISSME[18]	44.4	20.7	ICCV2015
ReRank[9]	77.1	63.6	CVPR2017
OIM Loss[78]	82.1	60.9	CVPR2017
MSCAN[25]	76.3	53.1	CVPR2017
DCA[25]	80.3	57.5	CVPR2017
DCGAN[8]	78.0	56.2	ICCV2017
k-reciprocal[9]	77.1	63.6	CVPR2017
OL-MANS[82]	60.7	–	ICCV2017
SVDNet[83]	82.3	62.1	ICCV2017
PA[84]	81.0	63.4	ICCV2017
JLML[79]	85.1	65.5	IJCAI2017
DSR[80]	82.7	61.2	CVPR2018
DeformGAN[81]	80.6	61.3	CVPR2018
Pose-transfer[26]	79.8	58.0	CVPR2018
FMN[13]	86.0	67.1	PRL2019
Ours	**86.5**	**76.2**	

the-art methods on both Rank-1 and mAP, comparing to the prevalent methods, such as online instance matching (OIM) Loss[78], k-reciprocal[9], joint learning multi-loss (JLML)[79], and deep spatial feature construction (DSR)[80], although they devote to design complicated networks architecture or various loss for better performance. Furthermore, our method outperforms other GAN based Re-ID methods including deep convolutional generative adversarial networks (DCGAN)[79], DeformGAN[81] and Pose-transfer[26], which aim to synthesize person images with various poses and image styles. This indicates that the t complementary advantages from different modalities play a more important role than the cross-view pose variation and style adaption in person Re-ID.

4.2.2 Comparison on DukeMTMC-reID dataset

The evaluation results of our method on DukeMTMC-reID dataset[8] is shown in Table 3. Our method consistently outperforms the state-of-the-art methods including either metric learning based methods, e.g., Bow+KISSME[18] and OIM Loss[78], or GAN based methods, e.g., DCGAN[8], similarity preserving generative adversarial networks (SPGAN)[85] and Pose-transfer[26]. Consistent with the results on Market1501[18], our method significantly improves the mAP metrics by 9 %, which verifies that our method can distinguish more challenging situations at the first rankings.

Table 3 Experimental comparison of the proposed approach with state-of-the-art methods on DukeMTMC-reID[8] (in %)

Methods	DukeMTMC-re-ID		
	Rank-1	mAP	References
Bow+KISSME[18]	25.1	12.2	ICCV2015
LOMO+XQDA[3]	30.8	17.0	CVPR2015
OIM Loss[78]	68.1	47.4	CVPR2017
SVDNet[83]	67.6	45.8	ICCV2017
DCGAN[8]	67.7	47.1	ICCV2017
Verif+Identif[84]	68.9	49.3	ICCV2017
SPGAN[85]	41.1	22.3	CVPR2018
Pose-transfer[26]	68.6	48.0	CVPR2018
Ours	**69.2**	**55.0**	

4.2.3 Comparison on CUHK03 dataset

Table 4 shows the results of our method with the state-of-the-art methods on the CUHK03 dataset[6]. Note that CUHK03[6] contains two folds, one of which is named as detected, where the person images/bounding boxes are obtained by pedestrian detector, while the other one by handcraft named as labeled. We test our method on the handcraft one labeled comparing it to the state-of-the-art methods. Our method achieves promising performance with 87.6 % and 84.1 % on Rank-1 and mAP respectively. It seems that our method has not improved as much as on the other two datasets Market1501[18] and DukeMTMC-reID[8]. The main reason is the limited challenges in the CUHK03 dataset[6], which contains few background clutters and illumination changes. In other words, our method can better improve the person Re-ID performance in more challenging scenarios.

Table 4 Experimental comparison of the proposed approach with state-of-the-art methods on CUHK03[6] (in %)

Methods	CUHK03		
	Rank-1	mAP	References
OIM Loss[78]	44.4	20.7	CVPR2017
MSCAN[25]	–	74.2	CVPR2017
DCA[25]	–	74.2	CVPR2017
PA[84]	85.4	–	ICCV2017
OL-MANS[82]	–	61.7	ICCV2017
JLML[79]	83.2	–	IJCAI2017
k-reciprocal[9]	61.6	67.6	CVPR2017
DCSL[86]	80.2	–	IJCAI2016
Deep pyramidal feature learning (DPFL)[11]	86.7	83.8	ICCV2017
Ours	**87.6**	**84.1**	

4.3 Qualitative examples

Fig. 5 (a) demonstrates two ranking results of corresponding queries on the Market1501 dataset[18], CUHK03 dataset[6] and DukeMTMC-reID dataset[8] respectively. Benefitting from the thermal information generated from the RGB modality, our method can overcome the challenges of background clutter (especially for Query (i), (ii) and (vi)), pose changes (especially for Query (i) to (iv) and (vi)), occlusions (especially for Query (iv) and (v)), and huge illumination changes (especially for Query (ii) and (vi)).

4.4 Ablation study

To verify the contribution of each component in our method, we evaluate several variants on the three datasets in this section, as shown Fig. 6. It is clear to see that: 1) By introducing the channel and spatial attention (CSA) or the TIR generation module, we can improve the rank-1 and especially the mAP accuracies, which verify the contributions of both components. 2) By integrating both CSA and TIR generation modules, we can further boost the performance which verifies the effectiveness of the proposed method. 3) TIR generation contributes more on Market1501[18] while CSA plays a more important role on the other two datasets DukeMTMC-reID[8] and CUHK03[6]. The reason is that the images of Market1501[18] consist of different background and occlusion and the complexity of the dataset is high, and TIR data could reduce the impact of these factors. The attention network can select discriminative regions of images

(a) (b)

Fig. 5 Qualitative examples of the proposed method; (a) Ranking results of the proposed method on three benchmark person re-identification datasets, where the left column indicates the query images, and the following ten columns are the corresponding top-10 hits obtained by our method; (b) Ranking results of the proposed method comparing with the baseline on Market1501 dataset[18]. The green and the red boxes indicate the right and the wrong hits respectively.

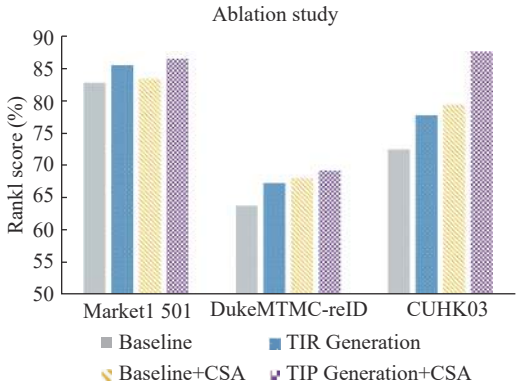


Fig. 6 Ablation study of the variants of our method on Market1501[18], DukeMTMC-reID[8] and CUHK03[6]

and important channels of feature maps for different modal data in three datasets.

Fig. 5 (b) illustrates our method is better than the baseline especially in some challenges such as background clutter, pose changes on Market1501[18]. We also compare our CSA module with the widely used SE-block[87] based on ResNet-50 on Market1501. As shown in Table 5, our attention module outperforms the SE-block.

4.5 Evaluation on backbones

To evaluate the generality of our method, we further evaluate our method with different backbones, including ResNet-101[75], ResNet-34[75], Res2Net-50[88] and SeNet[87], besides ResNet-50[75]. Table 6 reports the results of our method with various backbones. Our CSA module is more suitable to the ResNet network and could achieve better performance. It is clear to see that, our method achieves promising performance on all the backbones.

Table 5 Evaluation on CSA attention module comparing with SE-block[87] on Market1501[18]

Module	Rank-1	mAP
+CSA	83.4	61.0
+CSA+TIR generation	86.5	76.2
+SE-block	81.8	60.1
+SE-block+TIR generation	82.0	71.2

Furthermore, our TIR generation module and CSA module can boost the performance on each backbone, which verifies the contribution of the proposed method.

ResNet-101[75] slightly outperforms ResNet-50[75] by deeper convolution layers which could extract more high-level features. The remaining three backbones are overshadowed since Res2Net-50 and SeResNet-50 contain a large number of parameters leading to overfitting and more computation, ResNet-34 is shallow network which performs generally.

4.6 Parameter analysis

The important parameter in our method is α in (4), which balances the weight of RGB and thermal modalities during Re-ID as shown in Fig. 7. The larger α, the higher contribution of thermal information. We analyze the impact of α by varying 0.2, 0.5, 0.8, 1.0, 1.2, 1.5, 2.0, and observe that: 1) Our method with different weights consistently outperforms the baseline, which validates the effectiveness of generated TIR information. 2) Our method achieves the best performance in the range $0.8 < \alpha < 1.2$, which indicates that the generated TIR information contributes more or less the same as the RGB information. 3) A larger weight on TIR information may decline

Table 6 Evaluation on backbones with various components on Market1501[18], DukeMTMC-reID[8] and CUHK03[6]. The top three results are highlighted in red, green and blue, respectively

Backbones	Methods	Market1501		DukeMTMC-reID		CUHK03	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
ResNet-50	Baseline	82.8	59.9	63.8	43.0	72.5	66.0
	+ CSA	83.4	61.0	68.0	47.6	79.4	72.5
	+ TIR generation	85.5	66.0	67.3	47.1	77.8	74.2
	+ TIR generation + CSA	86.5	76.2	69.2	55.0	87.6	84.1
ResNet-101	Baseline	83.5	61.9	68.7	48.9	80.5	75.5
	+ CSA	84.8	64.2	69.6	50.3	84.2	77.3
	+ TIR generation	86.4	68.8	69.9	49.4	82.8	79.5
	+ TIR generation + CSA	87.4	78.2	70.6	54.2	87.8	84.3
ResNet-34	Baseline	75.4	51.4	55.8	32.9	63.0	56.7
	+ CSA	78.2	54.3	57.3	34.2	70.2	60.8
	+ TIR generation	79.0	56.8	60.1	37.6	68.5	63.2
	+ TIR generation + CSA	82.5	72.3	61.0	45.6	78.8	74.4
Res2Net-50	Baseline	76.0	61.2	56.2	33.6	63.6	57.7
	+ CSA	78.2	65.5	58.1	36.0	71.1	61.9
	+ TIR generation	76.8	62.1	60.3	37.9	70.5	65.1
	+ TIR generation + CSA	81.2	68.6	62.1	46.8	80.2	75.3
SeResNet-50	Baseline	79.1	66.8	65.1	45.3	69.0	63.7
	+ CSA	81.1	68.2	66.5	46.1	75.2	67.8
	+ TIR generation	80.1	67.2	66.1	45.8	72.5	65.2
	+ TIR generation + CSA	82.0	71.2	67.8	52.3	82.3	77.0

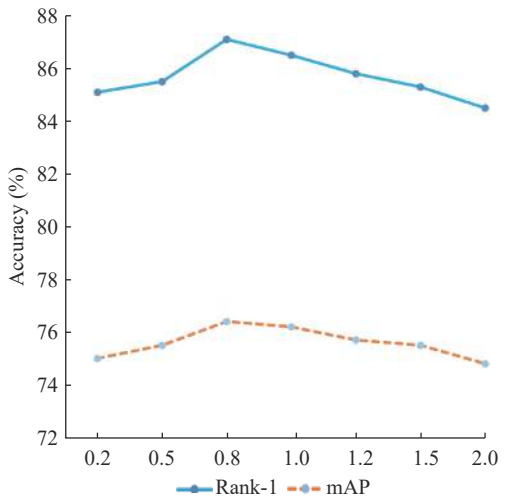


Fig. 7 Evaluation with different weights of the generated thermal data on Market-1 501 dataset[18]

the overall performance due to less appearance information in TIR data compared to RGB data. In this work, we set α to 1 to balance the channel weights of the generated thermal and visible in the input. Higher α, lower illumination and more background clutters.

5 Conclusions

In this work, we have proposed a RGBT representation learning network for person re-identification. It utilizes the generated model to obtain TIR data to solve hard backgrounds in Re-ID datasets. Benefiting from the thermal modality, it can learn more discriminative feature representation with both RGB and synthesised TIR information for person Re-ID. Furthermore, we have utilized a cross-modal attention network to adaptively integrate the multi-modal information for Re-ID. Our proposed framework achieves state-of-the-art performance on person Re-ID without additional computational cost. In the future, we will investigate more modality information to improve the robustness of single RGB modality based Re-ID tasks.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Nos. 61976002, 61976003 and 61860206004), Natural Science Foundation of Anhui Higher Education Institutions of China (No. KJ2019A0033), and the Open Project Program of the National Laboratory of Pattern Recognition (No.

201900046).

References

[1] O. Oreifej, R. Mehran, M. Shah. Human identity recognition in aerial images. In *Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, pp. 709–716, 2010. DOI: 10.1109/CVPR.2010.5540147.

[2] A. Mignon, F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, Providence, USA, pp. 2666–2672. 2012. DOI: 10.1109/CVPR.2012.6247987.

[3] S. C. Liao, Y. Hu, X. Y. Zhu, S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 2197–2206, 2015. DOI: 10.1109/CVPR.2015.7298832.

[4] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, H. Bischof. Large scale metric learning from equivalence constraints. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Providence, USA, pp. 2288–2295, 2012. DOI: 10.1109/CVPR.2012.6247939.

[5] A. X. Li, K. X. Zhang, L. W. Wang. Zero-shot fine-grained classification by deep feature learning with semantics. *International Journal of Automation and Computing*, vol. 16, no. 5, pp. 563–574, 2019. DOI: 10.1007/s11633-019-1177-8.

[6] W. Li, R. Zhao, T. Xiao, X. G. Wang. DeepReID: Deep filter pairing neural network for person re-identification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, pp. 152–159, 2014. DOI: 10.1109/CVPR.2014.27.

[7] L. Chen, H. Yang, S. Wu, Z. Y. Gao. Data generation for improving person re-identification. In *Proceedings of the 25th ACM International Conference on Multimedia*, ACM, MountainView, USA, pp. 609–617, 2017. DOI: 10.1145/3123266.3123302.

[8] Z. D. Zheng, L. Zheng, Y. Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 3774–3782, 2017. DOI: 10.1109/ICCV.2017.405.

[9] Z. Zhong, L. Zheng, D. L. Cao, S. Z. Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 3652–3661, 2017. DOI: 10.1109/CVPR.2017.389.

[10] J. Satake, M. Chiba, J. Miura. Visual person identification using a distance-dependent appearance model for a person following robot. *International Journal of Automation and Computing*, vol. 10, no. 5, pp. 438–446, 2013. DOI: 10.1007/s11633-013-0740-y.

[11] Y. B. Chen, X. T. Zhu, S. G. Gong. Person re-identification by deep learning multi-scale representations. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, Venice, Italy, pp. 2590–2600, 2017. DOI: 10.1109/ICCVW.2017.304.

[12] Z. D. Zheng, L. Zheng, Y. Yang. Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 3037–3045, 2019. DOI: 10.1109/TCSVT.2018.2873599.

[13] G. D. Ding, S. Khan, Z. M. Tang, F. Porikli. Feature mask network for person re-identification. *Pattern Recognition Letters*, vol. 137, pp. 91–98, 2020. DOI: 10.1016/j.patrec.2019.02.015.

[14] L. Wu, R. C. Hong, Y. Wang, M. Wang. Cross-entropy adversarial view adaptation for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2081–2092, 2020. DOI: 10.1109/TCSVT.2019.2909549.

[15] D. S. Xu, J. Chen, C. Liang, Z. Wang, R. M. Hu. Cross-view identical part area alignment for person re-identification. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, pp. 2462–2466, 2019. DOI: 10.1109/ICASSP.2019.8683137.

[16] L. Wei, Z. Y. Wei, Z. M. Jin, Z. X. Yu, J. Q. Huang, D. Cai, X. F. He, X. S. Hua. SIF: Self-inspirited feature learning for person re-identification. *IEEE Transactions on Image Processing*, vol. 29, pp. 4942–4951, 2020. DOI: 10.1109/TIP.2020.2975712.

[17] D. Yi, Z. Lei, S. C. Liao, S. Z. Li. Deep metric learning for person re-identification. In *Proceedings of the 22nd International Conference on Pattern Recognition*, IEEE, Stockholm, Sweden, pp. 34–39, 2014. DOI: 10.1109/ICPR.2014.16.

[18] L. Zheng, L. Y. Shen, L. Tian, S. J. Wang, J. D. Wang, Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 1116–1124, 2015. DOI: 10.1109/ICCV.2015.133.

[19] X. B. Chang, T. M. Hospedales, T. Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 2109–2118, 2018. DOI: 10.1109/CVPR.2018.00225.

[20] J. J. You, A. C. Wu, X. Li, W. S. Zheng. Top-push video-based person re-identification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1345–1353, 2016. DOI: 10.1109/CVPR.2016.150.

[21] A. Hermans, L. Beyer, B. Leibe. In defense of the triplet loss for person re-identification, [Online], Available: https://arxiv.org/abs/1703.07737, 2017.

[22] J. Wang, Z. Wang, C. Liang, C. X. Gao, N. Sang. Equidistance constrained metric learning for person re-identification. *Pattern Recognition*, vol. 74, pp. 38–51, 2018. DOI: 10.1016/j.patcog.2017.09.014.

[23] X. K. Zhu, X. Y. Jing, F. Zhang, X. Y. Zhang, X. G. You, X. Cui. Distance learning by mining hard and easy negative samples for person re-identification. *Pattern Recognition*, vol. 95, pp. 211–222, 2019. DOI: 10.1016/j.patcog.2019.06.007.

[24] H. T. Yao, S. L. Zhang, R. C. Hong, Y. D. Zhang, C. S. Xu, Q. Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2860–2871, 2019. DOI: 10.1109/TIP.2019.2891888.

[25] D. W. Li, X. T. Chen, Z. Zhang, K. Q. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 7398–7407, 2017. DOI: 10.1109/CVPR.2017.782.

[26] J. X. Liu, B. B. Ni, Y. C. Yan, P. Zhou, S. Cheng, J. G. Hu. Pose transferrable person re-identification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 4099–4108, 2018. DOI: 10.1109/CVPR.2018.00431.

[27] Y. X. Ge, Z. W. Li, H. Y. Zhao, G. J. Yin, S. Yi, X. G. Wang, H. S. Li. FD-GAN: Pose-guided feature distilling GAN for robust person re-identification. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 1230–1241, 2018.

[28] Z. D. Zheng, X. D. Yang, Z. D. Yu, L. Zheng, Y. Yang, J. Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 2133–2142, 2019. DOI: 10.1109/CVPR.2019.00224.

[29] T. Sattrupai, W. Kusakunniran. Deep trajectory based gait recognition for human re-identification. In *Proceedings of IEEE Region 10 Conference*, Jeju, South Korea, pp. 1723–1726, 2018. DOI: 10.1109/TENCON.2018.8650523.

[30] C. Carley, E. Ristani, C. Tomasi. Person re-identification from gait using an autocorrelation network. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Long Beach, USA, pp. 2345–2353, 2019. DOI: 10.1109/CVPRW.2019.00288.

[31] C. L. Li, X. Y. Liang, Y. J. Lu, N. Zhao, J. Tang. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition*, vol. 96, Article number 106977, 2019. DOI: 10.1016/j.patcog.2019.106977.

[32] C. L. Li, H. Cheng, S. Y. Hu, X. B. Liu, J. Tang, L. Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5743–5756, 2016. DOI: 10.1109/TIP.2016.2614135.

[33] L. St-Laurent, X. Maldague, D. Prevost. Combination of colour and thermal sensors for enhanced object detection. In *Proceedings of the 10th International Conference on Information Fusion*, IEEE, Quebec, Canada, pp. 1–8, 2007. DOI: 10.1109/ICIF.2007.4408003.

[34] D. T. Nguyen, H. G. Hong, K. W. Kim, K. R. Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, vol. 17, no. 3, Article number 605, 2017. DOI: 10.3390/s17030605.

[35] M. Ye, Z. Wang, X. Y. Lan, P. C. Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI, Stockholm, Sweden, pp. 1092–1099, 2018. DOI: 10.24963/ijcai.2018/152.

[36] P. Y. Dai, R. R. Ji, H. B. Wang, Q. Wu, Y. Y. Huang. Cross-modality person re-identification with generative adversarial training. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI, Stockholm, Sweden, pp. 677–683, 2018.

[37] M. Ye, X. Y. Lan, J. W. Li, P. C. Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI, New Orleans, USA, 2018.

[38] L. C. Zhang, A. Gonzalez-Garcia, J. van de Weijer, M. Danelljan, F. S. Khan. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1837–1850, 2019. DOI: 10.1109/TIP.2018.2879249.

[39] J. Y. Zhu, T. Park, P. Isola, A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 2242–2251, 2017. DOI: 10.1109/ICCV.2017.244.

[40] X. Zhang, Q. Yang. Transfer hierarchical attention network for generative dialog system. *International Journal of Automation and Computing*, vol. 16, no. 6, pp. 720–736, 2019. DOI: 10.1007/s11633-019-1200-0.

[41] B. S. Wang, G. Cao, Y. F. Shang, L. C. Zhou, Y. Q. Zhang, X. S. Li. Single-column CNN for crowd counting with pixel-wise attention mechanism. *Neural Computing and Applications*, vol. 32, no. 7, pp. 2897–2908, 2020. DOI: 10.1007/s00521-018-3810-9.

[42] T. V. Nguyen, Z. Song, S. Y. Yan. STAP: Spatial-temporal attention-aware pooling for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 1, pp. 77–86, 2015. DOI: 10.1109/TCSVT.2014.2333151.

[43] Z. Ji, K. L. Xiong, Y. W. Pang, X. L. Li. Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1709–1717, 2020. DOI: 10.1109/TCSVT.2019.2904996.

[44] Z. C. Wang, L. Du, F. Wang, H. T. Su, Y. Zhou. Multi-scale target detection in SAR image based on visual attention model. In *Proceedings of the IEEE 5th Asia-Pacific Conference on Synthetic Aperture Radar*, Singapore, Singapore, pp. 704–709, 2015. DOI: 10.1109/APSAR.2015.7306303.

[45] S. Woo, J. Park, J. Y. Lee, I. S. Kweon. CBAM: Convolutional block attention module. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 3–19, 2018. DOI: 10.1007/978-3-030-01234-2_1.

[46] H. R. Chen, Y. W. Wang, Y. M. Shi, K. Yan, M. Y. Geng, Y. H. Tian, T. Xiang. Deep transfer learning for person re-identification. In *Proceedings of the 4th International Conference on Multimedia Big Data*, IEEE, Xi'an, China, pp. 1–5, 2018. DOI: 10.1109/BigMM.2018.8499067.

[47] H. Y. Zhao, M. Q. Tian, S. Y. Sun, J. Shao, J. J. Yan, S. Yi, X. G. Wang, X. H. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 907–915, 2017. DOI: 10.1109/CVPR.2017.103.

[48] C. Su, J. N. Li, S. L. Zhang, J. L. Xing, W. Gao, Q. Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 3980–3989, 2017. DOI: 10.1109/ICCV.2017.427.

[49] Y. F. Sun, L. Zheng, Y. Yang, Q. Tian, S. J. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 501–518, 2018. DOI: 10.1007/978-3-030-01225-0_30.

[50] W. H. Chen, X. T. Chen, J. G. Zhang, K. Q. Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 1320–1329, 2017. DOI: 10.1109/CVPR.2017.145.

[51] Y. Yuan, W. Y. Chen, Y. Yang, Z. Y. Wang. In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Seattle, USA, pp. 1454–1463, 2020. DOI: 10.1109/CVPRW50498.2020.00185.

[52] I. B. Barbosa, M. Cristani, A. del Bue, L. Bazzani, V. Murino. Re-identification with RGB-D sensors. In *Proceedings of European Conference on Computer Vision*, Springer, Florence, Italy, pp. 433–442, 2012. DOI: 10.1007/978-3-642-33863-2_43.

[53] M. Munaro, A. Fossati, A. Basso, E. Menegatti, L. van Gool. One-shot person re-identification with a consumer depth camera. *Person Re-Identification*, S. G. Gong, M. Cristani, S. C. Yan, C. C. Loy, Eds., London, UK: Springer, pp. 161–181, 2014. DOI: 10.1007/978-1-4471-6296-4_8.

[54] F. Pala, R. Satta, G. Fumera, F. Roli. Multimodal person

reidentification using RGB-D cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 4, pp. 788–799, 2016. DOI: 10.1109/TCSVT.2015.2424056.

[55] A. Mogelmose, C. Bahnsen, T. Moeslund, A. Clapes, S. Escalera. Tri-modal person re-identification with RGB, depth and thermal features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Portland, USA, pp. 301–307, 2013. DOI: 10.1109/CVPRW.2013.52.

[56] X. X. Xu, W. Li, D. Xu. Distance metric learning using privileged information for face verification and person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 12, pp. 3150–3162, 2015. DOI: 10.1109/TNNLS.2015.2405574.

[57] V. John, G. Englebienne, B. Krose. Person re-identification using height-based gait in colour depth camera. In *Proceedings of IEEE International Conference on Image Processing*, Melbourne, Australia, pp. 3345–3349, 2013. DOI: 10.1109/ICIP.2013.6738689.

[58] A. C. Wu, W. S. Zheng, J. H. Lai. Robust depth-based person re-identification. *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2588–2603, 2017. DOI: 10.1109/TIP.2017.2675201.

[59] M. Paolanti, L. Romeo, D. Liciotti, R. Pietrini, A. Cenci, E. Frontoni, P. Zingaretti. Person re-identification with RGB-D camera in top-view configuration through multiple nearest neighbor classifiers and neighborhood component features selection. *Sensors*, vol. 18, no. 10, Article number 3471, 2018. DOI: 10.3390/s18103471.

[60] L. L. Ren, J. W. Lu, J. J. Feng, J. Zhou. Uniform and variational deep learning for RGB-D object recognition and person re-identification. *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4970–4983, 2019. DOI: 10.1109/TIP.2019.2915655.

[61] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS, Long Beach, USA, pp. 2672–2680, 2014.

[62] M. Mirza, S. Osindero. Conditional generative adversarial nets, [Online], Available: https://arxiv.org/abs/1411.1784, 2014.

[63] A. Radford, L. Metz, S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, [Online], Available: https://arxiv.org/abs/1511.06434, 2015.

[64] G. Perarnau, J. van de Weijer, B. Raducanu, J. M. Álvarez. Invertible conditional GANS for image editing, [Online], Available: https://arxiv.org/abs/1611.06355, 2016.

[65] P. Isola, J. Y. Zhu, T. H. Zhou, A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 5967–5976, 2017. DOI: 10.1109/CVPR.2017.632.

[66] D. Xu, W. L. Ouyang, E. Ricci, X. G. Wang, N. Sebe. Learning cross-modal deep representations for robust pedestrian detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 4236–4244, 2017. DOI: 10.1109/CVPR.2017.451.

[67] Y. Luo, J. Ren, M. Lin, J. H. Pang, W. X. Sun, H. S. Li, L. Lin. Single view stereo matching. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 155–163, 2018. DOI: 10.1109/CVPR.2018.00024.

[68] T. T. Qiao, J. Zhang, D. Q. Xu, D. C. Tao. MirrorGAN: Learning text-to-image generation by redescription. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 1505–1514, 2019. DOI: 10.1109/CVPR.2019.00160.

[69] L. Chen, S. Srivastava, Z. Y. Duan, C. L. Xu. Deep cross-modal audio-visual generation. In *Proceedings of Thematic Workshops of ACM Multimedia 2017*, ACM, Mountain View, USA, pp. 349–357, 2017. DOI: 10.1145/3126686.3126723.

[70] H. Zhou, Y. Liu, Z. W. Liu, P. Luo, X. G. Wang. Talking face generation by adversarially disentangled audio-visual representation. *In Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 9299–9306, 2019. DOI: 10.1609/aaai.v33i01.33019299.

[71] C. Li, M. Wand. Precomputed real-time texture synthesis with Markovian generative adversarial networks. In *Proceeding of 4th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 702–716, 2016. DOI: 10.1007/978-3-319-46487-9_43.

[72] A. C. Wu, W. S. Zheng, H. X. Yu, S. G. Gong, J. H. Lai. RGB-infrared cross-modality person re-identification. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 5390–5399, 2017. DOI: 10.1109/ICCV.2017.575.

[73] D. P. Kingma, J. Ba. Adam: A method for stochastic optimization. [Online], Available: https://arxiv.org/abs/1412.6980, 2014.

[74] B. T. Zhang, X. P. Wang, Y. Shen, T. Lei. Dual-modal physiological feature fusion-based sleep recognition using CFS and RF algorithm. *International Journal of Automation and Computing*, vol. 16, no. 3, pp. 286–296, 2019. DOI: 10.1007/s11633-019-1171-1.

[75] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[76] S. Ioffe, C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, pp. 448–456, 2015.

[77] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 17–35, 2016. DOI: 10.1007/978-3-319-48881-3_2.

[78] T. Xiao, S. Li, B. C. Wang, L. Lin, X. G. Wang. Joint detection and identification feature learning for person search. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 3376–3385, 2017. DOI: 10.1109/CVPR.2017.360.

[79] W. Li, X. T. Zhu, S. G. Gong. Person re-identification by deep joint learning of multi-loss classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, Melbourne, Australia, pp. 2194–2200, 2017. DOI: 10.24963/ijcai.2017/305.

[80] L. X. He, J. Liang, H. Q. Li, Z. N. Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 7073–7082, 2018. DOI: 10.1109/CVPR.2018.00739.

[81] A. Siarohin, E. Sangineto, S. Lathuilière, N. Sebe. Deformable GANs for pose-based human image generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 3408–3416, 2018. DOI: 10.1109/CVPR.2018.00359.

[82] J. H. Zhou, P. Yu, W. Tang, Y. Wu. Efficient online local

metric adaptation via negative samples for person re-identification. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 2439–2447, 2017. DOI: 10.1109/ICCV.2017.265.

[83] Y. F. Sun, L. Zheng, W. J. Deng, S. J. Wang. SVDNet for pedestrian retrieval. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 3820–3828, 2017. DOI: 10.1109/ICCV.2017.410.

[84] L. M. Zhao, X. Li, Y. T. Zhuang, J. D. Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 3239–3248, 2017. DOI: 10.1109/ICCV.2017.349.

[85] W. J. Deng, L. Zheng, Q. X. Ye, G. L. Kang, Y. Yang, J. B. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 994–1003, 2018. DOI: 10.1109/CVPR.2018.00110.

[86] Y. Q. Zhang, X. Li, L. M. Zhao, Z. F. Zhang. Semantics-aware deep correspondence structure learning for robust person re-identification. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, New York, USA, pp. 3545–3551, 2016.

[87] J. Hu, L. Shen, G. Sun. Squeeze-and-excitation networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 7132–7141, 2018. DOI: 10.1109/CVPR.2018.00745.

[88] S. H. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, M. H. Yang, P. H. S. Torr. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. DOI: 10.1109/TPAMI.2019.2938758.

Ai-Hua Zheng received the B. Eng. and Ph. D. degrees in computer science and technology from Anhui University, China in 2006 and 2008, respectively. And she received the Ph. D. degree in computer science from University of Greenwich, UK in 2012. She visited University of Stirling and Texas State University from June to September in 2013 and from September 2019 to August 2020, respectively. She is currently an associate professor and Ph. D. supervisor in School of Computer Science and Technology in Anhui University, China. As the first author or corresponding author, she has published more than 40 academic papers, including top conferences papers in *American Association for Artificial Intelligence Conference on Artificial Intelligence* (AAAI) and the *International Joint Conference on Artificial Intelligence* (IJCAI), and authoritative journals in *IEEE Transactions on Systems, Man, and Cybernetics: Systems* (TSMCS), *Pattern Recognition* (PR), *Pattern Recognition Letters* (PRL), *Neurocomputing* (NeuCom), *Cognitive Computation* (CogCom), the *IEEE International Symposium on Network Computing and Applications* (NCA), etc. She is a member of China Computer Federation (CCF) and China Society of Image and Graphics (CSIG). She is also serving as reviewers for representative conferences and journals, including AAAI, IJCAI, *IEEE Transactions on Image Processing* (TIP), *IEEE Transactions on Multimedia* (TMM), *IEEE Transactions on Intelligent Transportation Systems* (TITS), PR, etc. She has obtained the Best Paper Award in the *International Conference on Software Engineering Research, Management and Applications* (SERA) 2017 and the Best Student Paper Award in the workshop in the *IEEE International Conference on Multimedia and Expo* (ICME) 2019.

Her research interests include vision based artificial intelligence and pattern recognition, especially on person/vehicle re-identification, audio visual computing, and multi-modal intelligence.

E-mail: ahzheng214@foxmail.com
ORCID iD: 0000-0002-9820-4743

Zi-Han Chen received the B. Eng. degree in software engineering from Anhui University, China in 2018. He is currently a master student in computer science and technology from Anhui University, China.

His research interests include computer vision, person re-identification and machine learning.

E-mail: zhchen96@stu.ahu.edu.cn
ORCID iD: 0000-0002-5991-5462

Cheng-Long Li received the M. Sc. and Ph. D. degrees in computer science from School of Computer Science and Technology, Anhui University, China in 2013 and 2016, respectively. From 2014 to 2015, he worked as a visiting student with School of Data and Computer Science, Sun Yat-sen University, China. He was a postdoctoral research fellow at the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China. He is currently an associate professor at School of Computer Science and Technology, Anhui University, China. He was a recipient of the ACM Hefei Doctoral Dissertation Award in 2016.

His research interests include computer vision and deep learning.

E-mail: lcl1314@foxmail.com (Corresponding author)
ORCID iD: 0000-0002-7233-2739

Jin Tang received the B. Eng. degree in automation and the Ph. D. degree in computer science from Anhui University, China in 1999 and 2007, respectively. He is a professor with School of Computer Science and Technology, Anhui University.

His research interests include computer vision, pattern recognition and machine learning.

E-mail: tangjin@ahu.edu.cn

Bin Luo received the B. Eng. degree in electronics, and the M. Eng. degree in computer science from Anhui University, China in 1984 and 1991, respectively, and the Ph. D. degree in computer science from University of York, UK in 2002. From 2000 to 2004, he was a research associate with University of York, UK. He is currently a professor with Anhui University, China.

His research interests include graph spectral analysis, large image database retrieval, image and graph matching, statistical pattern recognition, digital watermarking and information security.

E-mail: luobin@ahu.edu.cn