Deep Attention Network for RGB-Infrared Cross-Modality Person Re-Identification

Yang Li¹,²,* and Huahu Xu²

¹ School of Information Technology, Shanghai Jianqiao University, Shanghai 201306, China
² School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China

*E-mail: 16083@gench.edu.cn

Abstract. RGB-Infrared cross-modality person re-identification is an important task for 24-hour full-time intelligent video surveillance, the task is challenging because of cross modal heterogeneity and intra modal variation. A novel deep attention network is proposed in this paper to handle these challenges by increasing the discriminability of the learned person features. The method includes three elements: (1) dual-path CNN to extract the feature maps of the RGB images and infrared images respectively, (2) dual-attention mechanism combining spatial attention and channel attention to enhance the discriminability of extracted features, and (3) joint loss function joining bi-directional ranking loss and identity loss to constraint the training process to further increase the accuracy. Extensive experiments on two public datasets demonstrate the effectiveness of our proposed method because the method achieves higher performance than state-of-the-arts methods.

1. Introduction

Person re-identification (reID) aims to solve the problem of person retrieval in the non-overlapping camera network, because its importance in video surveillance, it has attracted the continuous attention of researchers in the field of computer vision. At present, most existing methods mainly study person reID under visible light image (RGB image) [1,2]. However, in a 24-hour intelligent surveillance system, it is not enough to only use visible light image. For example, at night, because of poor or no light, the visible light camera can not capture enough person appearance information. In this case, it is very necessary for image acquisition equipment that does not rely on visible light, such as infrared camera. Infrared cameras are widely used in intelligent surveillance system to capture person appearance information according to the infrared radiation of human body. Now many cameras use not only visible light imaging, but also infrared imaging function, which can obtain person images in daytime and night scenes through the same camera, so as to achieve full-time intelligent surveillance.

In a 24-hour intelligent surveillance system, the person image to be probed may be the visible light image taken in the daytime, while the gallery images to be retrieved may be the infrared image taken in the night, and vice versa. It means the probe images and the gallery images are from different modalities. Therefore, it is necessary to study the cross-modality person reID. This paper will focus on the problem of RGB-infrared person reID, which is also referred as visible-thermal person reID.

For the task of RGB-infrared cross-modality person reID, there are two major issues that need to be handled: cross modal heterogeneity and intra modal variation. On the one hand, the cross modal
heterogeneity is the most critical problem in RGB-infrared person reID. Cross modal heterogeneity mainly refers to the huge differences in the images of the same person acquired by different image sensors. For example, visible RGB images vs. infrared grayscale images are mainly caused by the difference of perception between different waveband spectra of visible and infrared cameras. There are three channels of RGB in visible image, which contain a lot of color information of persons, while the infrared image has only one channel and contains some infrared information. This will make it difficult for color information to be used in this kind of RGB-infrared person reID task based on heterogeneous data. However, color information is the most important information to distinguish different persons in visible light person reID. On the other hand, as with the traditional visible light person reID, the person changes inside modality caused by different person poses, changing lighting conditions, occlusion, background clutter and perspective differences still exist, which will increase the difficulty of RGB-infrared person reID and make it a more challenging task.

To deal with the two key problems of the task, cross-modality heterogeneity and intra-modal variation, this paper attempts to enhance the discriminative feature learning from the following aspects:

- **Dual-path convolutional neutral network (CNN) network structure.** As in [3], dual-path CNN network structure is used to extract person features, including one visible light network and one infrared network. First, two independent CNN networks are used to extract modal specific information, which is used to emphasize the differences between different modalities. Then, some parameters sharing full-connected network layers are used to embed the modal specific information into a common space. Such a dual-path network structure can extract the cross modal sharing features, which simultaneously considers the extraction of different and common information between multiple modalities.

- **Dual-attention mechanism.** In order to further enhance the discrimination of features, spatial attention and channel attention are introduced to explore visual feature dependencies in the spatial and channel dimensions, respectively. For spatial attention, a non-local attention mechanism [4] is introduced to utilize the spatial dependencies between two locations in the feature map to better capture the fine-grained pixel-level saliency. For channel attention, an improved SE Network [5] is introduced to capture the channel dependencies between two channel maps. The re-weighted channel maps are merged together to increase the sensitivity of feature representations. The two attentions are combined to further enhance the discrimination of feature representations.

- **Joint loss function.** The joint loss function includes bi-directional ranking loss and identity loss. Bi-directional ranking loss is introduced to address the large inter-class variations of both cross-modality and intra-modality by making the distance of furthest positive sample pairs to be smaller than the nearest negative sample pairs with predefined margin. Identity loss is introduced to address the large intra-class variations by treating the same person identity cross-modality as the same class and making the learnt features are identity invariant.

2. **Methodology**

A novel deep attention network framework is proposed to learn the discriminative feature representations for RGB-infrared cross-modality person reID task, the framework is shown in figure 1. The network consists of three parts: (1) Dual-path CNN backbone network, (2) dual-attention module (DAM) composed of channel attention (CA) and spatial attention (SA), and (3) joint loss function, including bi-directional ranking loss and identity loss. Two ResNet50 [6] are used as the backbone of each network path, and the two ResNet50 are independent of each other, which is used to extract modal independent information. Then the attention module further explores the spatial and channel dependence of the features extracted by ResNet50 to obtain more discriminative feature representations, and then the high-level network adopt the full-connection (FC) layers which sharing parameters across modalities to connect the output of the two attention modules, mapping the multi-modal person information to a common space. Finally, the network is trained under the guidance of bi-directional ranking loss and identity loss. It should be noted that the full-connection (FC) layers in the
dotted box is parameter shared across two paths/modalities.

**Figure 1.** Framework of proposed deep attention network.

### 2.1. Dual-path CNN

A dual-path CNN structure is used to extract person features, including a visible light network and an infrared network. In each network, any CNN model designed for image classification can be used as the backbone network. In this paper, ResNet50 is used because of its excellent performance in many person reID methods [7,8], as well as its relatively simple structure. The dual-path ResNet50 backbone networks are independent of each other, aiming to learn the information of specific modalities to emphasize the problem of cross modal heterogeneity.

As shown in figure 1, parameters sharing full-connection layers (FC2 and FC3) are used to map the specific modal information into a public space, in order to learn the person features shared by multi-modalities and reduce the differences between heterogeneous modalities. The specific settings are as follows. (1) Global average pooling (GAP) is connected the output feature map of attention module. (2) Next, a full-connection layer (FC1) is used as a bottleneck to reduce the dimension. (3) Then a batch normalization layer (BN) and a full-connection layer (FC2), its output will be used to for the bi-directional ranking loss. (4) Finally, another full-connection layer (FC3) is used to perform classification under softmax identity loss. The output dimension of FC3 is the number of classes, which corresponds to the number of different person identities in the datasets.

### 2.2. Dual-attention module

A dual-attention module (DAM) is introduced to further enhance the discrimination of features, the module includes spatial attention and channel attention.

#### 2.2.1. Channel attention (CA)

For an input image, each channel in the feature map can be regarded as a feature detector, and channel attention pays attention to which types of features are considered to be more meaningful. SE network [5] is used for the channel attention in this paper. Channel attention is applied to the output of ResNet50 to explore the inter-channel correlation, as shown in figure 2.

**Figure 2.** Architecture of channel attention.

The spatial information of each channel in the feature map \( S \in R^{C \times W \times H} \) is aggregated through
GAP layer to generate the channel feature descriptor \( s = [s_1, s_2, \ldots, s_c] \in \mathbb{R}^{C \times 1 \times 1} \), and the channel attention map \( m \) can be determined as follows:

\[
m = \text{sigmoid} \left( W_2 \text{ReLU} \left( W_1 \left( \sum_{w,h} (S_c)_{w,h} \right) \right) \right)
\]

where \( W_1 \) and \( W_2 \) refer to the weight matrix of the fully connected layer FC1 and FC2.

Finally, the refined feature map \( S' \in \mathbb{R}^{C \times W \times H} \) is calculated by multiplying the channel attention map with the original input feature map \( m \):

\[
S' = S \otimes m
\]

where \( \otimes \) is channel-wise multiplication. The network uses channel attention map \( m \) to dynamically mine the interdependencies between channel maps to learn meaningful feature maps and suppress less useful feature maps, which helps to improve the discriminability of feature representation.

2.2.2. Spatial attention (SA). Spatial attention is utilized to capture the rich contextual relevance and enhance the ability of feature representation by modeling the context information in feature map. As shown in figure 3, the output of the channel attention module is first fed into the convolution layer to generate the feature map \( \in \mathbb{R}^{C \times W \times H} \), reducing the dimension from 2048 to 512. Then three new feature maps \( \{X, Y, Z\} \in \mathbb{R}^{C \times D} \) are generated through 1 \( \times \) 1 convolution, respectively. And they are reshaped as \( \mathbb{R}^{C \times D} \), where \( D = W \times H \). Then, we multiply the transposition of \( X \) and \( Y \), and use the softmax to calculate the spatial attention matrix \( S \in \mathbb{R}^{D \times D} \):

\[
s_{ij} = \frac{\exp(X_i \cdot Y_j)}{\sum_{i=1}^{D} \exp(X_i \cdot Y_j)}
\]

where \( s_{ij} \) means the influence of position \( i \) on position \( j \). Then the matrix multiplication is performed between the feature map \( Z \) and the transposition of the spatial attention matrix \( S \), and the result is reshaped to \( \mathbb{R}^{C \times W \times H} \). Finally, the result is multiplied by the parameter \( \alpha \), then the element sum operation is performed with the feature map \( I \), and the final output \( S' \in \mathbb{R}^{W \times H \times C} \) is:

\[
S' = \alpha S \cdot Z + I
\]

It can be seen from equation (4) that the final feature of each position is related to the features of all positions. Therefore, it contains global context information and selectively captures spatial dependencies among positions based on spatial attention. The spatial attention improves the spatial mutual contribution and robustness of the task.

2.2.3. Combining channel attention and spatial attention. Channel attention and spatial attention can produce complementary attention features by focusing on the types and positions of effective features respectively. Therefore, they can be used together. There are three arrangements of the combination of the two attentions: sequential in CA-SA order, sequential in SA-CA order and in parallel. As shown in figure 1, we arranged the two attentions sequentially in CA-SA order because it can achieve better
performance than other two arrangements.

2.3. Joint loss function

2.3.1. Bi-directional ranking loss. Given query person image set \( G = \{x_i|0 \leq i < N, z_i|0 \leq i < N\} \), which contains \( N \) RGB person images and \( N \) infrared person images. Define the RGB image \( x_i \) to be matched, and its label \( y_i \), when using the sample \( x_i \) to match in the gallery set \( G \), we hope that: (1) the images matching with the sample \( x_i \) in \( G \) are on top of the mismatched images. (2) The images with higher matching scores are on top of the images with lower matching scores. In RGB-infrared cross-modality person reID, for a RGB image \( x_i \), ranking loss is to make the distance between the RGB image \( x_i \) and the positive sample infrared image \( z_j \), less than the distance between the RGB image \( x_i \) and the negative sample infrared image \( z_k \). The formula is as follows:

\[
D(x_i, z_j) < D(x_i, z_k) - \beta, \forall y_i = y_j, \forall y_i \neq y_k
\]

where \( \beta \) is pre-defined margin. Ranking loss can be obtained after transform above formula as follows:

\[
L_{\text{rank}} = \max[\beta + D(x_i, z_j) - D(x_i, z_k), 0], \forall y_i = y_j, \forall y_i \neq y_k
\]

For cross-modality task, ranking loss actually includes bi-directional relationship: triplet loss from RGB to infrared (one RGB image, two positive and negative infrared images) and triplet loss from infrared to RGB (one infrared image, two positive and negative RGB images), which can be defined as bi-directional ranking loss. The bi-directional ranking loss can be expressed by below formula:

\[
L_{\text{bi-rank}} = \sum_{\forall y_i = y_j, \forall y_i \neq y_k} \max[\beta + D(x_i, z_j) - D(x_i, z_k), 0],
\]

\[
+ \sum_{\forall y_i = y_j, \forall y_i \neq y_k} \max[\beta + D(z_i, x_j) - D(z_i, x_k), 0]
\]

where the subscripts \( i \) and \( j \) have the same person identity, \( i \) and \( k \) have different person identity, \( L_2 \) norm is used to normalize all the input eigenvectors of \( x \) and \( z \), Euclidean distance is used to calculate the distance between samples.

2.3.2. Identity loss. Identity loss is often used in the person reID for visible light images [9]. In the cross-modality person reID, existing researches also use person identity information to supervise the training process [10]. Identity loss does not consider the differences across modalities, only focuses on the person identity information. Identity loss can be expressed as:

\[
L_{\text{id}} = -\frac{1}{m} \sum_{i=1}^{m} \log \frac{e^{w_i^T x_i + b_i}}{\sum_{j=1}^{n} e^{w_j^T x_i + b_j}}
\]

where \( w_{y_i} \) is the \( i \)-th weight of the last full-connection layer, \( b \) is the bias of the full-connection layer, \( y_i \) is the identity label of the input person sample \( x_i \), \( m \) is the batch size, and \( n \) is the total number of person identity.

2.3.3. Joint loss function. Finally, the above two losses are weighted and summed, and the final loss function can be defined as:

\[
L = L_{\text{bi-rank}} + \lambda L_{\text{id}}
\]

where \( \lambda \) the weight coefficients to adjust the importance of the two losses, which need to be determined by experiments.
3. Experiments

3.1. Experimental settings

3.1.1. Datasets. The experiments are based on two datasets: SYSU-MM01 [11] and RegDB [12].

SYSU-MM01 is a large-scale dataset collected by six cameras, including four visible light cameras and two infrared cameras. The training set includes 395 persons, including 22258 RGB images and 11909 infrared images. The test set consists of 96 persons, including 3803 infrared images as probe images and 301 randomly selected RGB images as gallery images.

RegDB is collected by dual camera systems, and contains 412 persons. It contains 10 RGB images and 10 infrared images for each person. The dataset is randomly split into two halves for training and testing. During testing, the probe images and gallery images should come from different modality. The procedure is repeated 10 trials to achieve statistic results.

3.1.2. Evaluation metrics. The cumulative matching characteristics (CMC) and the mean average precision (mAP) are adopted to evaluate the performances for fair comparison since the two metrics are commonly used for RGB-infrared cross-modality person reID task.

3.1.3. Implementation details. The implementation is based on PyTorch open source machine learning library with Python 3.6. The ResNet50 network for feature extraction uses the weights pre-trained on ImageNet to initialize, and SGD is used to optimize the loss function, the momentum is set to 0.9, and the decay weight is set to 0.0005. During the training, the initial learning rate is set as 0.01, the total training epoch is 80, and the batch size is 32. The learning rate remains unchanged at 0.01 in the first 30 epochs, and then decayed by a coefficient of 0.1 between 30 and 60 epochs, and decayed by a coefficient of 0.01 after 60 epochs, so that the algorithm converges quickly in the early stage and adjusts slowly in the later stage.

3.2. Comparison with state-of-the-arts methods

We compared our method with some state-of-the-arts (SOTA) methods, including zero-padding [11], TONE [10], DCTR [3], cmGAN [13] and D^2RL [14]. The comparison results are shown in table 1.

| Methods               | RegDB | SYSU-MM01 |
|-----------------------|-------|-----------|
|                       | r = 1 | r = 10    | r = 20    | mAP   | r = 1 | r = 10 | r = 20    | mAP   |
| zero-padding [11]     | 17.75 | 34.21    | 44.35    | 18.90 | 14.80 | 54.12 | 71.33    | 15.95 |
| TONE [10]             | 16.87 | 34.03    | 44.10    | 14.92 | 12.52 | 50.72 | 68.60    | 14.42 |
| TONE + XQDA [10]      | 21.94 | 45.05    | 55.73    | 21.80 | 14.01 | 52.78 | 69.06    | 15.97 |
| TONE + HCML [10]      | 24.44 | 47.53    | 56.78    | 20.80 | 14.32 | 53.16 | 69.17    | 16.16 |
| DCTR (BCTR) [3]       | 32.67 | 57.64    | 66.58    | 30.99 | 16.12 | 54.90 | 71.47    | 19.15 |
| DCTR (BDTR) [3]       | 33.47 | 58.42    | 67.52    | 31.83 | 17.01 | 55.43 | 71.96    | 19.66 |
| cmGAN [13]            | -     | -        | -        | -     | -     | 26.97 | 67.51    | 80.56 |
| D^2RL [14]            | 43.40 | 66.10    | 76.30    | 44.10 | 28.90 | 70.60 | 82.40    | 29.20 |
| Ours                  | 52.23 | 71.08    | 80.96    | 52.31 | 36.75 | 83.24 | 92.69    | 40.21 |

From table 1, it can be seen that the performance of the deep attention network method proposed in this paper on the two datasets RegDB and SYSU-MM01 is far better than the best SOTA method D^2RL [14], on the RegDB dataset, Rank1 is increased by 8.83, and mAP is increased by 8.21; on the SYSU-MM01 dataset, Rank1 is increased by 7.85, and mAP is increased by 10.31. This proves the effectiveness of the proposed deep attention network method, including the integration of dual attention mechanism and joint loss function (bi-directional ranking loss and identity loss), in enhancing the discriminability of person features for RGB-infrared cross-modality person reID.
4. Conclusion
Aiming at the problem that infrared images and visible light images have great differences in features, this paper proposes a deep attention network to deal with cross-modal and intra-modal differences by enhancing feature discrimination. The ResNet50 is used as the backbone network to extract the features of infrared image and RGB image. The dual attention mechanism including spatial attention and channel attention is proposed to enhance the discriminability of person features. The learning process is constrained by the joint loss function based on bi-directional ranking loss and identity loss to further improve the accuracy. The experimental results show that the deep attention network proposed in this paper is effective for the RGB-infrared cross-modality person re-identification task.

References
[1] Ye M, Shen J, Lin G, Xiang T, Shao L and Hoi S. Deep learning for person re-identification: A survey and outlook[J]. arXiv preprint arXiv:2001.04193, 2020.
[2] Li Y, Xu H and Bian M. End to end person re-identification based on attention mechanism[C]// IOP Conference Series: Materials Science and Engineering 646 (2019) 012053.
[3] Ye M, Wang Z, Lan X and Yuan P. Visible thermal person re-identification via dual-constrained top-ranking[C]//Proceedings of the International Joint Conferences on Artificial Intelligence, 2018: 1092-1099.
[4] Wang X, Girshick R, Gupta A and He K. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7794-7803.
[5] Hu J, Shen L and Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.
[6] He K., Zhang X, Ren S and Sun J. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
[7] Wang C, Zhang Q, Huang C, Liu, W and Wang X. Mancs: A multi-task attentional network with curriculum sampling for person re-identification[C]//Proceedings of the European Conference on Computer Vision. 2018: 384-400.
[8] Sun Y, Zheng L, Yang Y, Tian Q and Wang S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]//Proceedings of the European Conference on Computer Vision. 2018: 501-518.
[9] Wu D, Zheng S, Bao W, Zhang X, Yan C and Huang D. A novel deep model with multi-loss and efficient training for person re-identification[J]. Neurocomputing, 2019, 324: 69-75.
[10] Ye M, Lan X, Li J and Yuan P. Hierarchical discriminative learning for visible thermal person re-identification[C] // Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[11] Wu A, Zheng W, Yu H, Gong S and Lai J. RGB-infrared cross-modality person re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 5380-5389.
[12] Nguyen D, Hong H, Kim K and Park K. Person recognition system based on a combination of body images from visible light and thermal cameras[J]. Sensors, 2017, 17(3): 605.
[13] Dai P, Ji R, Wang H, Wu Q and Huang Y. Cross-modality person re-identification with generative adversarial training[C]//International Joint Conference on Artificial Intelligence. 2018: 677–683.
[14] Wang Z, Wang Z, Zheng Y, Chuang Y and Satoh S. Learning to reduce dual-level discrepancy for infrared-visible person re-identification[C]//Proceedings of the European Conference on Computer Vision. 2019: 618–626.