IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 22, 2019, accepted November 17, 2019, date of publication November 26, 2019,
date of current version December 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955930

Visible Infrared Cross-Modality Person Re-Identification Network Based on Adaptive Pedestrian Alignment

BO LI[1], **XIAOHONG WU**[1], **QIANG LIU**[1], **XIAOHAI HE**[1], **(Member, IEEE),**
AND FEI YANG[2], **(Member, IEEE)**
[1]College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China
[2]TAL AI Lab, Danling SOHO, Beijing 100080, China

Corresponding authors: Xiaohong Wu (wxh@scu.edu.cn) and Fei Yang (yang.fei@100tal.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61871278, in part by the Fund of Sichuan University - Tomorrow Advancing Life (TAL), in part by the Chengdu Science and Technology Project under Grant 2016-XT00-00015-GX, and in part by the Sichuan Science and Technology Program under Grant 2018HH0143.

ABSTRACT Cross-modality person re-identification between the visible domain and infrared domain is important but extremely challenging for night-time surveillance. Besides the cross-modality discrepancies caused by different camera spectrums, visible infrared person re-identification (VI-REID) still suffers from much pedestrian misalignment as well as the variations caused by different camera viewpoints and various pedestrian pose deformations like traditional person re-identification. In this paper, we propose a multi-path adaptive pedestrian alignment network (MAPAN) to learn discriminative feature representations. The multi-path network learns features directly from the data in an end-to-end manner and aligns the pedestrians adaptively without any additional manual annotations. To alleviate the intra-modality discrepancies caused by image misalignment, we combine the aligned visible image features with the original visible image features and enhance the attention of the network towards pedestrians, resulting in significant improvements in distinguishability of the learning features. To mitigate the cross-modality discrepancies between the visible domain and the infrared domain, the discriminative features of the two modalities are mapped to the same feature embedding space, and the identity loss as well as triplet loss is incorporated as the overall loss. Extensive experiments demonstrate the superior performance of proposed method compared to the state-of-the-arts.

INDEX TERMS Person re-identification, pedestrian alignment, visible infrared cross-modality, triplet loss.

I. INTRODUCTION

Person Re-identification (known as ReID) is a technique in the field of computer vision to identify a specific pedestrian as (numerically) the same particularly as one encountered on a previous occasion [1]. It is generally considered to be a sub-problem of image retrieval and has a bright application prospect in the field of intelligent monitoring. But there are great challenges for ReID such as low resolution of the camera and various pedestrian pose deformations. Pedestrian images captured by different cameras may also cause enormous discrepancies in the appearance of pedestrians due to occlusion, various viewpoints, illumination variations, etc.

Despite the difficulties, traditional person re-identification has made great progress in recent years with people's unremitting efforts, including many supervised methods [2]–[11], as well as unsupervised or weakly supervised methods [12]–[19]. Most of the existing methods are mainly based on feature representation learning [2], [5], [7], [20] or metric learning [3], [6], [10], [16]. Recently, the newly proposed methods tend to work on body part-based features and semantic information [8], [11], [17], [23], or attention mechanisms [9], [21], [22] to achieve higher recognition accuracy.

However, all the traditional person re-identification methods mentioned above only use visible images to match visible images whereas the visible camera can not capture clear images under poor illumination environments. Fortunately, with the development of society, most of the cameras are equipped with infrared camera function today. Infrared cameras can acquire infrared image information of pedestrians during day or night, which provides favorable conditions

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Sharif.

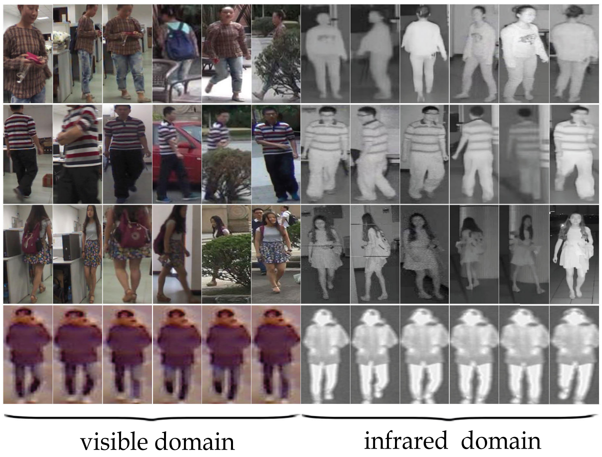


FIGURE 1. Some examples of visible images and infrared images in dataset SYSU-MM01 (the first three rows) and RegDB (the 4-th row) for VI-REID. Besides visual characteristic discrepancies caused by different camera spectrums, we can observe that the visible pedestrian images (particularly in dataset SYSU-MM01) are still more or less with occlusion, excessive background, scale variations, etc. So the task of cross-modality person re-identification between the visible domain and infrared domain (VI-REID) is extremely challenging.

for visible infrared cross-modality person re-identification (VI-REID). In recent years, a lot of researches [24]–[29] on cross-modality person re-identification between the visible domain and infrared domain have been conducted. Currently, VI-REID is the same as most cross-modality retrieval matching tasks. Different from the traditional person re-identification, VI-REID mainly focuses on matching cross-modality images. It usually uses visible(infrared) pedestrian images to search for the infrared(visible) pedestrian images across camera devices [26]. As shown in Fig.1, the challenges it faces are not only the problems of traditional person re-identification such as occlusion, illuminations variations and scale deviations, but also the problem of the discrepancies between the heterogeneous data [25]. The three-channel information of visible images is different from the single-channel information of infrared images in information capacity and representation. Different resolutions and lighting conditions may have different effects on the two types of images. For example, applying the same lighting condition to both types of images may increase the contrast of visible images, whereas for infrared images it may be too bright to be clear. Some work [27]–[29] attempts to improve pedestrian matching accuracy by reducing the cross-modality discrepancies of the heterogeneous data. In fact, there are also lots of intra-modality discrepancies caused by image misalignment, especially inside the visible images. As shown in Fig.1 and Fig.2, there are plenty of images with excessive backgrounds, occluded images, incomplete images in the cross-modality dataset, particularly in SYSU-MM01 [25]. These image misalignment phenomena are usually caused by unsatisfactory camera capturing angles and image post-processing errors during dataset acquisition.

Inspired by the space transformer network [30], we propose a multi-path adaptive pedestrian alignment network(MAPAN) strategy to deal with the intra-modality discrepancies caused by image misalignment. At the same time, in order to mitigate the enormous differences between the heterogeneous data, the visible image features and infrared image features extracted by ResNet50 are mapped to the same feature embedding space, and the identity loss as well as triplet loss is incorporated as the overall loss. The main contributions are summarized as follows:

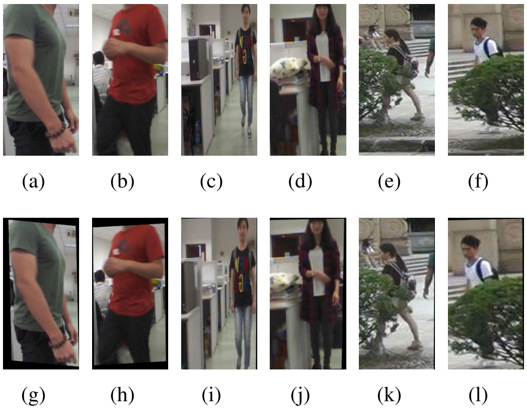


FIGURE 2. Image misalignment in dataset SYSU-MM01. We mainly show the image misalignment in SYSU-MM01 because there are fewer misalignment pedestrians in the RegDB dataset. (a,b) are incomplete pedestrian images; (c,d) are pedestrian images with excessive background; (e,f) are occluded pedestrian images. (g-l) are the corresponding visible pedestrian images aligned by the proposed MAPAN. We can notice that the pedestrian alignment is not perfect and obvious, but it more or less reduces the scale variations and position deviations.

- We propose an end-to-end multi-path adaptive pedestrian alignment network(MAPAN) strategy to deal with the intra-modality discrepancies in misaligned images caused by the acquisition of the cross-modality dataset for the first time.
- We map the visible image features and infrared image features to the same feature embedding space, and combine the identity loss and triplet loss as the overall loss, alleviating the discrepancies between heterogeneous data effectively, and achieve superior experimental performance compared to the state-of-the-arts.

The rest of this paper is organized as follows: We describe the related work in the next section. The proposed method based on multi-path adaptive pedestrian alignment network is presented in Section III. Section IV includes experiments and analysis and section V is conclusion.

II. RELATED WORK

A detailed overview about traditional person re-identification in visible domain can be found in [1], [31], and the overview of other asymmetric cross-modality person re-identifications is introduced in [32], [33] in detail. Here we mainly discuss the visible infrared cross-modality person re-identification.

Recently, [24] proposed a person re-identification method based on the combination of the pedestrian visible images and infrared images captured by a visible light camera and a infrared camera respectively, and disclosed the pedestrian dataset RegDB containing both visible images and infrared

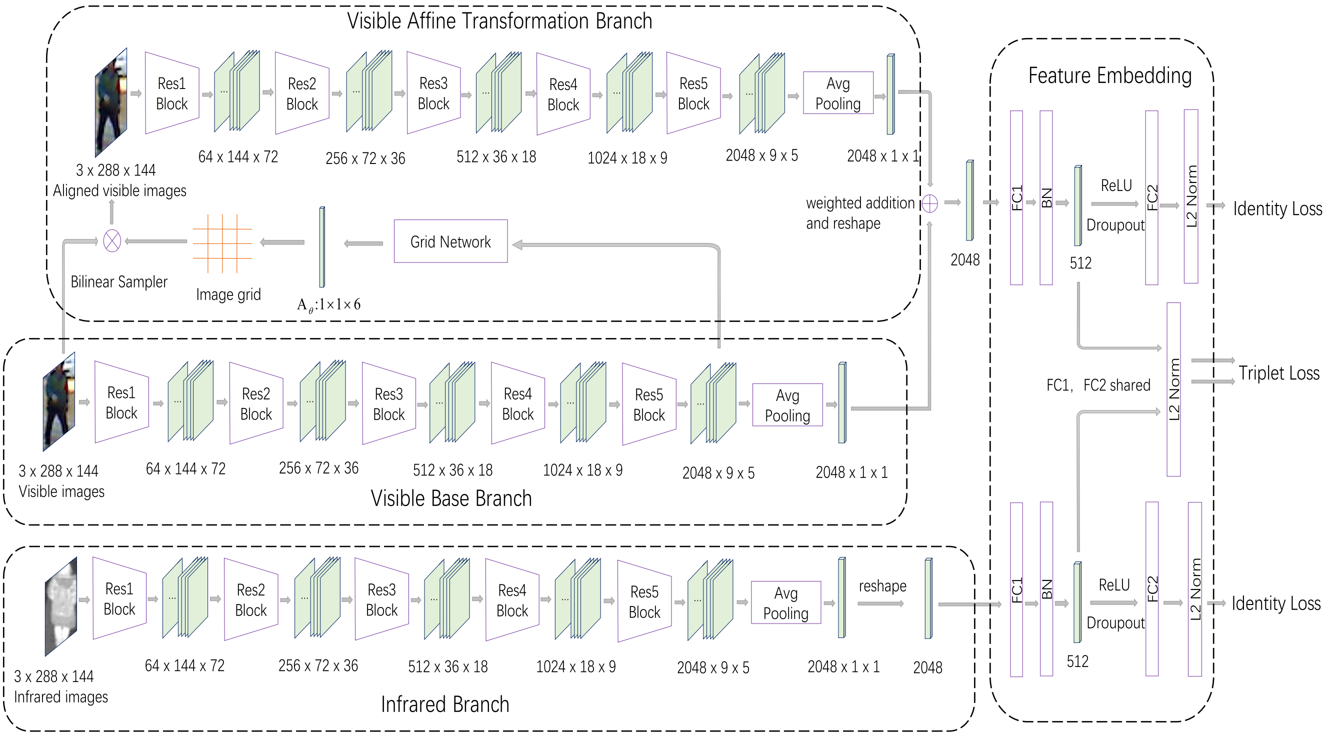


FIGURE 3. **Illustrations of proposed MAPAN. It comprises two main components: multi-path network for feature extraction and fully connected layers for feature embedding. Specifically, the multi-path network comprises three branches: infrared branch, visible base branch and visible affine transformation branch.**

images. Because the additional modality information captured by infrared camera was integrated with standard RGB visible images, the person re-identification performance was improved efficiently. It was the first time that infrared images had been adopted for person re-identification.

Subsequently, [25] raised the visible infrared cross-modality re-identification(VI-REID) problem for the first time and contributed a large scale cross-modality pedestrian dataset SYSU-MM01 for VI-REID, and a deep zero-padding method was proposed that utilizes a one-stream network to capture information for a specific domain. For VI-REID, the lack of authentication information to re-identify the same person between visible domain and infrared domain, and the difficulty to learn a robust representation for such a large-scale cross-modality person retrieval are the two main challenges. [28] proposed a novel cross-modality generative adversarial network (termed as cmGAN) to tackle the challenges and achieved superior performance.

Existing VI-REID methods mainly focus on the cross-modality discrepancies caused by the heterogeneous data, whereas VI-REID also suffers from the intra-modality discrepancies caused by the different camera viewpoints, pedestrian pose variations and deformations, making the mixed discrepancies more serious. Therefore, [27] proposed a hierarchical cross-modality person matching model by optimizing the modality-shared and modality-specific metrics learning jointly. Reference [26] proposed a dual-path network combining the identity loss and a novel bi-directional dual-constrained top-ranking loss to learn discriminative representations of visible domain and infrared domain, and the proposed dual-path network with a novel loss constrained the cross-modality and intra-modality discrepancies simultaneously and mitigated cross-modality discrepancies effectively. Reference [29] put forward that previous VI-REID methods usually only considered the feature-level constraints to optimize the model. However it was difficult to handle the mixed discrepancies mentioned above without image-level constraints. In [29], a novel dual level discrepancy reduction learning scheme was proposed to handle the cross-modality and intra-modality discrepancies by utilizing image-level and feature-level constraints separately.

None of the above methods focus on the problem of image misalignment inside the dataset. As shown in Fig.1 and Fig.2, there are plenty of visible pedestrian images with excessive background, occluded images and incomplete pedestrian images in the large scale dataset SYSU-MM01. In the RegDB dataset, image misalignment is relatively less serious. In order to achieve pedestrian alignment and enhance attention of the network towards pedestrians, we design a multi-path learning framework with adaptive affine transformation structure without any human intervention, and an identity loss and a batch hardest triplet loss are incorporated to the framework to handle large cross-modality and intra-modality discrepancies.

III. PROPOSED METHOD

This paper proposes a multi-path end-to-end feature learning framework MAPAN for VI-REID as shown in Fig.3. The framework learns the feature representations and distance metrics in an end-to-end manner while preserving high discriminability. It comprises two main components: multi-path network for feature extraction and fully connected layers for feature embedding. Specifically, the multi-path network comprises three branches: the visible affine transformation branch, visible base branch and infrared branch, all of them do not share weights. The visible base branch is identical to the infrared branch structure, and both of them use the residual network ResNet-50 [34], including 5 down-sampled blocks and an average pooling layer. The visible affine transformation branch consists of a grid network, a bilinear sampler and a residual network ResNet-50. As mentioned above, there are more misalignment phenomena in the input visible images, and in order to obtain more robust visible features by affine transformation correction, we fuse the feature of visible base branch and the feature of affine transformation branch by weighted addition.

A. INFRARED BRANCH AND VISIBLE BASE BRANCH

Both the infrared branch and the visible branch inputs are three-channel infrared images and visible images with a height and width of 288 × 144, respectively. We assume that $\mathcal{X}(\mathcal{Z})$ denote a batch of input visible(infrared) images. In the infrared branch, the features extracted for the infrared images are represented by $\phi_{\mathcal{I}}(\mathcal{Z})$, and the features extracted by the visible base branch and by the visible affine transformation branch are represented by $\phi_{\mathcal{V}}(\mathcal{X})$ and $\phi'_{\mathcal{V}}(\mathcal{X}')$ respectively, where $\mathcal{X}'$ denote transformed images generated by the affine transformation of $\mathcal{X}$ in the visible affine transformation branch.

B. VISIBLE AFFINE TRANSFORMATION BRANCH

The visible affine transformation branch consists of a bilinear sampler, a grid network, and a residual network ResNet50 [34] with the same structure as visible base branch. As shown in Fig.3, the bilinear sampler takes the batch of input visible images $\mathcal{X}$ as inputs. The grid network contains an average pooling layer and two fully connected layers, and takes the fifth residual block features extracted from the visible base branch as inputs.

The high-level feature map contains the low-level feature map of the original image and reflects the local pattern information [35]–[37]. As shown in Fig.5, it is obvious that the visible base branch's high responses and attentions are mostly concentrated on the pedestrian bodies, no matter whether the images are aligned or not. Therefore, we can feed the feature map of the fifth residual block into the grid network to regress a set of 6-dimensional transformation parameter $\mathcal{A}_\theta$, which is used to guide the affine transformation to align the pedestrians. Specifically, the learned transformation parameter $\mathcal{A}_\theta$ is used to generate an image grid for the bilinear sampler, and the point-by-point conversion process from the target images to the source images is formulated as:

$$\begin{pmatrix} x_k^s \\ y_k^s \end{pmatrix} = \mathcal{A}_\theta \begin{pmatrix} x_k^t \\ y_k^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_k^t \\ y_k^t \\ 1 \end{pmatrix} \tag{1}$$

where (x_k^t, y_k^t) is the k-th target coordinate in the regular grid of the transformed images, (x_k^s, y_k^s) is the source coordinate of the sampled point in the input images, and $\mathcal{A}_\theta$ is the affine transformation matrix, where $\theta_{11}, \theta_{12}, \theta_{21}$ and θ_{22} mainly control the size of the transformed images and rotation changes while θ_{13} and θ_{23} control the offset of the transformed images. Note that the coordinate mapping is mapped from the target images to the input images. Since the transformation matrix $\mathcal{A}_\theta$ contains continuous differentiable decimals and the target coordinate (x_k^t, y_k^t) is discrete, the source coordinate (x_k^s, y_k^s) will be continuous. So when we find the correspondence between the target coordinates and the source coordinates by Equation (1), a certain sampling strategy is required to generate the transformed images. Here, we use the frequently-used bilinear sampling, then the correspondence between the input images $\mathcal{X}$ and the output images $\mathcal{X}'$ is formulated as:

$$\mathcal{X}'_{ij} = \sum_{n=1}^{\mathcal{H}} \sum_{m=1}^{\mathcal{W}} \mathcal{X}_{nm}\, [1-|x_k^s-m|\,]_+ \, [1-|y_k^s-n|\,]_+$$
$$i \in [1\ldots,\mathcal{H}], \quad j \in [1\ldots,\mathcal{W}], \; k \in [1\ldots,\mathcal{WH}] \tag{2}$$

where $\mathcal{X}'_{ij}$ represents the pixel value of the coordinate (i, j) position in each channel in the target images, $\mathcal{X}_{nm}$ represents the pixel value of each channel in the source images at coordinate (n, m), $\mathcal{H}$ and $\mathcal{W}$ denote the height and width of the input images respectively, and $[\zeta]_+$ denotes *max* $(\zeta, 0)$. The bilinear sampling is continuous and steerable, so the above equation is steerable and allows loss gradient back propagation. Its partial derivatives with respect to $\mathcal{X}_{nm}$ and x_k^s are:

$$\frac{\partial \mathcal{X}'_{ij}}{\partial \mathcal{X}_{nm}} = \sum_{n=1}^{\mathcal{H}} \sum_{m=1}^{\mathcal{W}} \left[1-|x_k^s-m|\,\right]_+ \left[1-|y_k^s-n|\,\right]_+ \tag{3}$$

$$\frac{\partial \mathcal{X}'_{ij}}{\partial x_k^s} = \sum_{n=1}^{\mathcal{H}} \sum_{m=1}^{\mathcal{W}} \begin{cases} 0 & |x_k^s - m| \geq 1 \\ \mathcal{X}_{nm}\, [1-|y_k^s-n|\,]_+, & x_k^s < m \\ -\mathcal{X}_{nm}\, [1-|y_k^s-n|\,]_+, & x_k^s > m \end{cases} \tag{4}$$

Obviously, $\frac{\partial \mathcal{X}'_{ij}}{\partial y_k^s}$ is similar to $\frac{\partial \mathcal{X}'_{ij}}{\partial x_k^s}$. After the above affine transformation generating $\mathcal{X}'$, the features $\phi'_{\mathcal{V}}(\mathcal{X}')$ are extracted by the residual network of the visible affine transformation branch.

C. THE FEATURE EMBEDDING

Through the above three branches, we can get the aforementioned features $\phi_{\mathcal{I}}(\mathcal{Z})$, $\phi_{\mathcal{V}}(\mathcal{X})$, $\phi'_{\mathcal{V}}(\mathcal{X}')$. Virtually, both $\phi_{\mathcal{V}}(\mathcal{X})$ and $\phi'_{\mathcal{V}}(\mathcal{X}')$ are features extracted from $\mathcal{X}$, in order to make full use of these two features, we choose to fuse $\phi'_{\mathcal{V}}(\mathcal{X}')$ and $\phi_{\mathcal{V}}(\mathcal{X})$ by weighted addition, that is, $\lambda\phi_{\mathcal{V}}(\mathcal{X}) + (1-\lambda)\phi'_{\mathcal{V}}(\mathcal{X}')$ are the final features extracted from $\mathcal{X}$, where λ is a predefined trade-off parameter ranging from 0 to 1 to balance the contributions of the two features. We will later demonstrate the complementarity between $\phi_{\mathcal{V}}(\mathcal{X})$ and $\phi'_{\mathcal{V}}(\mathcal{X}')$ through experiments so as to illustrate the rationality of weighted fusion of these two features. Thus, we will only consider the distance metric of the features $\phi_{\mathcal{I}}(\mathcal{Z})$ of the input infrared images and the fused features $\lambda\phi_{\mathcal{V}}(\mathcal{X}) + (1-\lambda)\phi'_{\mathcal{V}}(\mathcal{X}')$ of the input visible images. Since our image retrieval task essentially matches the visible images with the infrared images, it is necessary to map the features of the visible images and the infrared images to the same feature space to reduce cross-modality differences between the infrared images and the visible images. So we map $\phi_{\mathcal{I}}(\mathcal{Z})$ and $\lambda\phi_{\mathcal{V}}(\mathcal{X}) + (1-\lambda)\phi'_{\mathcal{V}}(\mathcal{X}')$ to the same feature space with the function $f_\theta(\cdot)$ parameterized by θ, which is exactly a linear transformation, to output embedding features $f_\theta(\phi_{\mathcal{I}}(\mathcal{Z}))$ and $f_\theta(\lambda\phi_{\mathcal{V}}(\mathcal{X}) + (1-\lambda)\phi'_{\mathcal{V}}(\mathcal{X}'))$. For clarity, We simply refer to $f_\theta(\phi_{\mathcal{I}}(\mathcal{Z})$ and $f_\theta(\lambda\phi_{\mathcal{V}}(\mathcal{X}) + (1-\lambda)\phi'_{\mathcal{V}}(\mathcal{X}'))$ as $f_\theta(\mathcal{Z})$ and $f_\theta(\mathcal{X})$.

D. THE OVERALL LOSS

We use the conventional cross entropy loss to predict pedestrian identities. As described in the experiment section IV, our sampling strategy is that in each batch, for a dataset containing $\mathcal{N}$ identities, $\mathcal{P}$ identities are randomly selected first, and then for each identity, we randomly acquire $\mathcal{K}$ visible pedestrian images and $\mathcal{K}$ infrared pedestrian images, thus resulting in a batch of $2 \times \mathcal{P} \times \mathcal{K}$ images. The feature embedding fully connected layer $f_\theta(\cdot)$ outputs 512-dimensional feature tensors $f_\theta(\mathcal{Z})$ and $f_\theta(\mathcal{X})$. The next fully connected layer $f_\beta(\cdot)$ parameterized by β generates $\mathcal{N}$-dimensional feature tensors $f_\beta(f_\theta(\mathcal{Z}))$ and $f_\beta(f_\theta(\mathcal{X}))$. For convenience, we use $\mathcal{X}_j^i$ to represent the j-th image of i-th person(with an identity of i) in the batch $\mathcal{X}$, the same is true for $\mathcal{Z}$. Assuming that $\hat{p}_x^{i,j} = softmax(f_\beta(f_\theta(\mathcal{X}_j^i)))$ and $\hat{p}_z^{i,j} = softmax(f_\beta(f_\theta(\mathcal{Z}_j^i)))$, then $\hat{p}_x^{i,j}$ and $\hat{p}_z^{i,j}$ represent the identity predicted probabilities of input pedestrians $\mathcal{X}_j^i$ and $\mathcal{Z}_j^i$ respectively. For example, $\hat{p}_x^{i,j}(k)$ represents the predicted probability that the input visible image $\mathcal{X}_j^i$ has an identity of k. Given the true label $p_x^{i,j}$ and $p_z^{i,j}$ for $\mathcal{X}_j^i$ and $\mathcal{Z}_j^i$ with target identity of i, which means $p_x^{i,j}(i) = 1, p_x^{i,j}(k) = 0\ \forall k \neq i$ and $p_z^{i,j}(i) = 1, p_z^{i,j}(k) = 0\ \forall k \neq i$. Our batch identity loss is defined as follows:

$$\ell_{identity}(\theta; \mathcal{X}, \mathcal{Z}) = -\frac{1}{\mathcal{PKN}} \sum_{i=1}^{\mathcal{P}} \sum_{j=1}^{\mathcal{K}} (\sum_{k=1}^{\mathcal{N}} p_x^{i,j}(k) log(\hat{p}_z^{i,j}(k)) + \sum_{k=1}^{\mathcal{N}} p_z^{i,j}(k) log(\hat{p}_z^{i,j}(k))) \tag{5}$$

The identity loss $\ell_{identity}(\theta; \mathcal{X}, \mathcal{Z})$ only considers the identity of each input sample, and does not emphasize whether the input $\mathcal{X}$ and $\mathcal{Z}$ belong to the same identity or not. To further mitigate cross-modality variations between infrared images and visible images, we consider designing a batch hardest triplet loss [38] to optimize the metric embedding function $f_\theta(\cdot)$. For each infrared anchor sample $\mathcal{Z}_a^i$ in the batch, the core idea for computing triplet loss $\ell_{\mathcal{I}_{triplet}}(\theta; \mathcal{X}, \mathcal{Z})$ is that, we can select the hardest positive visible sample $\mathcal{X}_p^i$ with the same identity as $\mathcal{Z}_a^i$, whose embedding feature is furthest from $\mathcal{Z}_a^i$ in the feature space within the batch, and the hardest negative visible sample $\mathcal{X}_n^j (j \neq i)$ with different identity from $\mathcal{Z}_a^i$, whose embedding feature is nearest from $\mathcal{Z}_a^i$ in the feature space within the batch. And it is the same for each visible anchor sample $\mathcal{X}_a^i$ in the batch to compute $\ell_{\mathcal{V}_{triplet}}(\theta; \mathcal{X}, \mathcal{Z})$. $\ell_{\mathcal{I}_{triplet}}(\theta; \mathcal{X}, \mathcal{Z})$ and $\ell_{\mathcal{V}_{triplet}}(\theta; \mathcal{X}, \mathcal{Z})$ are formulated as follows respectively:

$$\ell_{\mathcal{I}_{triplet}}(\theta; \mathcal{X}, \mathcal{Z}) = \sum_{i=1}^{\mathcal{P}} \sum_{a=1}^{\mathcal{K}} [\max_{p=1,2...\mathcal{K}} \mathcal{D}(f_\theta(\mathcal{Z}_a^i), f_\theta(\mathcal{X}_p^i)) - \min_{\substack{j=1...\mathcal{P} \\ n=1...\mathcal{K} \\ j\neq i}} \mathcal{D}(f_\theta(\mathcal{Z}_a^i), f_\theta(\mathcal{X}_n^j)) + \xi]_+ \tag{6}$$

$$\ell_{\mathcal{V}_{triplet}}(\theta; \mathcal{X}, \mathcal{Z}) = \sum_{i=1}^{\mathcal{P}} \sum_{a=1}^{\mathcal{K}} [\max_{p=1,2...\mathcal{K}} \mathcal{D}(f_\theta(\mathcal{X}_a^i), f_\theta(\mathcal{Z}_p^i)) - \min_{\substack{j=1...\mathcal{P} \\ n=1...\mathcal{K} \\ j\neq i}} \mathcal{D}(f_\theta(\mathcal{X}_a^i), f_\theta(\mathcal{Z}_n^j)) + \xi]_+ \tag{7}$$

where ξ denotes a predefined positive threshold to control the minimum distance between positive and negative sample features. $\mathcal{D}(\cdot)$ represents the euclidean distance metric. Since we consider the cross-modality triplet loss $\ell_{\mathcal{I}_{triplet}}(\theta; \mathcal{X}, \mathcal{Z})$ as well as $\ell_{\mathcal{V}_{triplet}}(\theta; \mathcal{X}, \mathcal{Z})$ has equal effects on the optimization of the network, our overall loss $\ell_{overall}(\theta; \mathcal{X}, \mathcal{Z})$ is defined as:

$$\ell_{overall}(\theta; \mathcal{X}, \mathcal{Z}) = \beta\ell_{identity}(\theta; \mathcal{X}, \mathcal{Z}) + \alpha(\ell_{\mathcal{I}_{triplet}}(\theta; \mathcal{X}, \mathcal{Z}) + \ell_{\mathcal{V}_{triplet}}(\theta; \mathcal{X}, \mathcal{Z})) \tag{8}$$

where β and α are predefined trade-off parameters to balance the contributions of the identity loss and triplet loss.

IV. EXPERIMENTS

A. DATASETS AND SETTINGS

For VI-REID, there are two publicly available datasets till now: SYSU-MM01 [25] and RegDB [24]. They are all adopted for evaluation.

SYSU-MM01 [25] is a large-scale near-infrared(not thermal) cross-modality dataset which contains both visible images and infrared images. It's collected by four visible cameras(camera 1, 2, 4, and 5) and two infrared cameras(camera 3 and 6). The dataset is very challenging because some cameras are located in outdoor environments, and others are not. This dataset contains 491 different persons, and each person was captured by at least one visible camera and one infrared camera. We adopt the challenging single-shot all-search mode as evaluation protocol mentioned in [25].

There are 395 persons with 22258 visible images and 11909 infrared images in training set. For testing, there are 96 persons with 3803 infrared images for query and 301 pedestrian visible images are randomly selected as gallery set.

RegDB [24] is a far-infrared (thermal) cross-modality dataset, which is collaboratively collected by a visible camera and a infrared camera. It contains 412 persons, and each person has 10 visible images captured by the visible camera and 10 infrared images captured by the infrared camera. We adopt the valuation protocol mentioned in [27], where the RegDB dataset is randomly divided into two halves for training and testing respectively. For testing, the gallery set are the images from one modality while the query set are the images from the other modality.

In our experiments, the tests on the above two datasets were both repeated 10 trials to obtain statistically stable results.

B. IMPLEMENTATION DETAILS

1) EVALUATION METRICS

All experiments were evaluated by two commonly used evaluation metrics (the standard cumulated matching characteristics (CMC) and mean average precision (mAP)) with a single query setup.

2) BATCH SAMPLING STRATEGY

Since our overall loss combined with cross entropy loss and batch hardest triplet loss is slightly different from general person re-identifications, it's necessary to introduce the batch sampling strategy. As mentioned before, Specifically, $\mathcal{P}$ person identities are firstly randomly selected from $\mathcal{N}$ person identities for each iteration, where $\mathcal{N}$ is the total number of person identities. Then we randomly select $\mathcal{K}$ visible images and $\mathcal{K}$ infrared images of the selected identity to construct the batch, in which totally $2\times\mathcal{P}\times\mathcal{K}$ images are fed into the multi-path network for training. To calculate the visible infrared batch hardest triplet loss, we select $\mathcal{P}\times\mathcal{K}$ infrared images as anchors and corresponding hardest positive and hardest negative visible samples within the batch. Although the hardest samples are sampled within small subset of dataset at each iteration, the global optimum can be achieved after enough training iterations due to the mechanism of random sampling.

3) EXPERIMENTAL SETTINGS

We implement our algorithm with Pytorch framework and use ResNet-50 [34] as our pre-trained model. The training period epoch needs to be at least 60. Here we set both $\mathcal{P}$ and $\mathcal{K}$ to 6, thus the batch-size of single ResNet-50 [34] branch is 6 × 6, and so the input batch of the multi-branch network MAPAN contains 2 × 6 × 6 images: 36 visible images and 36 infrared images. The sizes of the two fully connected layers for feature embedding are set as 512 and the total number of person identities $\mathcal{N}$. The sizes of the two fully connected layers in grid network are set as 64 and 6. The learning rate is initially set to 0.01 for the fully connected feature embedding layers and 0.001 for the rest of MAPAN except for the grid network. Specifically, for the regression of the transformation parameter $\mathcal{A}_\theta$, the network is inclined to be stuck in a local minimum at the early iterations, so we use a relatively small learning rate 2×10^{-5} to stabilize the learning of parameter A_θ in the grid network. In addition,we initially set θ_{11} and θ_{22} to 0.85, and set $\theta_{12}, \theta_{13}, \theta_{21}, \theta_{23}$ to zero, which makes MAPAN's attention tend to concentrate on the center part of the visible images at early iterations, thus facilitating the network convergence, reducing time consumption and enhancing the MAPAN's attention towards pedestrian body areas. All of the learning rates are attenuated by 0.1 times for every 30 epochs except the learning rate for learning parameter $\mathcal{A}_\theta$ in grid network.

For the optimization of network parameters during training, we use the most widely used stochastic gradient descent(SGD) with a nesterov momentum fixed to 0.9 in general machine learning to achieve faster back propagation and smoother convergence. Followed by [38], the predefined positive threshold ξ for the triplet loss is set as $\xi = 1.2$. Due to the large difference in data distribution between the SYSU-MM01 dataset and the RegDB dataset, the overall loss we use for training on the two datasets is parameterized by different parameters. For the SYSU-MM01 dataset, we set the ratio of identity loss and triplet loss to 1:0.05, which means to set the trade-off parameters as $\beta = 1, \alpha = 0.05$. For the RegDB dataset, we set the ratio of identity loss and triplet loss to 1:1, which means to set the trade-off parameters as $\beta = 1, \alpha = 1$.

C. EXPERIMENT ANALYSIS

First, the input data is sampled according to the batch sampling strategy mentioned above, and resized to 288 × 144. Then we pad the input images on all sides with 10-pixel width zero values, resizing the size to 308×164. And after randomly cropped to 288×144, the sampled visible images and infrared images are horizontally flipped randomly with a probability of 0.5 and fed to the multi-path network for training, respectively. The visible pedestrian images are processed by visible base branch and visible affine transformation branch to output discriminative features. Then the visible features and the infrared features are fed to feature embedding layers and the loss gradients are back propagated. The proposed method not only effectively alleviates the misalignment problem of visible images, but also reduces the over-fitting of the network, and improves the robustness to some extent.

1) FEATURE FUSION STRATEGY

The fifth residual block features extracted by the visible base branch contain the position information of the pedestrian in the original images to some extent, as shown in Fig. 5. The grid network takes the fifth residual block features as input and outputs the affine transformation parameter $\mathcal{A}_\theta$. And then the input visible images $\mathcal{X}$ are affinely transformed with $\mathcal{A}_\theta$ to generate $\mathcal{X}'$. Finally the transformed visible images $\mathcal{X}'$ are fed to the residual network in the visible affine transformation branch to output the features $\phi'_{\mathcal{V}}(\mathcal{X}')$.

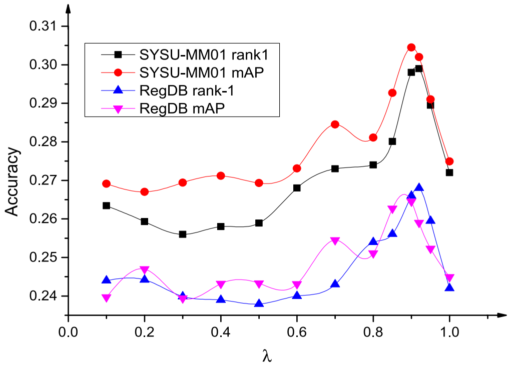


FIGURE 4. Performance of MAPAN with different λ in method A for dataset SYSU-MM01 and RegDB.

TABLE 1. Comparison of the fusion methods. Re-identification rates (%) at rank-r and mAP (%).

Method	rank-1	rank-5	rank-10	mAP
	SYSU-MM01			
A: $\lambda\phi_{\mathcal{V}}(\mathcal{X}) + (1-\lambda)\phi'_{\mathcal{V}}(\mathcal{X}')$	**29.80**	**57.43**	**72.47**	**30.45**
B: $[\phi_{\mathcal{V}}(\mathcal{X}), \phi'_{\mathcal{V}}(\mathcal{X}')]$	19.14	46.20	62.87	22.48
C: $\phi_{\mathcal{V}}(\mathcal{X})$	27.50	55.64	69.76	28.49
D: $\phi'_{\mathcal{V}}(\mathcal{X}')$	26.56	55.87	70.00	28.60
	RegDB			
A: $\lambda\phi_{\mathcal{V}}(\mathcal{X}) + (1-\lambda)\phi'_{\mathcal{V}}(\mathcal{X}')$	**26.60**	**40.43**	**49.15**	**26.37**
B: $[\phi_{\mathcal{V}}(\mathcal{X}), \phi'_{\mathcal{V}}(\mathcal{X}')]$	17.11	29.43	38.52	19.23
C: $\phi_{\mathcal{V}}(\mathcal{X})$	24.32	38.09	47.35	24.68
D: $\phi'_{\mathcal{V}}(\mathcal{X}')$	23.89	37.91	48.27	24.43

Both the $\phi_{\mathcal{V}}(\mathcal{X})$ extracted by the visible base branch and the $\phi'_{\mathcal{V}}(\mathcal{X}')$ extracted by the visible affine transform branch are features of $\mathcal{X}$. There are mainly two ways to fuse $\phi_{\mathcal{V}}(\mathcal{X})$ and $\phi'_{\mathcal{V}}(\mathcal{X})$ to obtain the final features f of the visible pedestrian images. One is feature weighted addition fusion: $f = \lambda\phi_{\mathcal{V}}(\mathcal{X}) + (1-\lambda)\phi'_{\mathcal{V}}(\mathcal{X}')$, and the other is the way of feature concatenating: $f = [\phi_{\mathcal{V}}(\mathcal{X}), \phi'_{\mathcal{V}}(\mathcal{X}')]$, where $[\cdot]$ denotes tensor concatenating. For these two fusion methods, we have done a lot of experimental explorations. Note that in order to facilitate the performance of the model, we use identity loss combined with triplet loss as our overall loss here. Table 1 is part of our experimental results, in which the method A represents feature weighted fusion and set the trade-off parameters as $\lambda = 0.9$, and the method B is the feature concatenating method, and the method C(D) is to directly use $f = \phi(\mathcal{X})$ or $f = \phi'(\mathcal{X}')$ as the final extraction features of visible images. It indicates that the method A, regardless of rank-1, rank-5 or mAP, outperforms the feature concatenating method B, and also outperforms the method C and D. Therefore, we deem that the feature weighted addition method A is more suitable for our task here. The following line graph Fig.4 are the results of our further explorations of the best value of λ. It can be observed that when $\lambda = 0.9$, rank-1 and map gets the maximum value almost simultaneously. In theory, we speculate that since $\phi(\mathcal{X})$ and $\phi'(\mathcal{X}')$ are essentially the same type of features, it is reasonable to obtain new fusion features in a way of weighted addition, and the feature concatenating in method B may achieve better results for essentially different features.

In addition, for the λ in method A, it denotes the proportion of the original visible features. Given a relatively small λ, the feature extractions of the visible original images will be supported by insufficient supervised information, resulting in some deviations of the transformation parameter $\mathcal{A}_\theta$ in regression. The extracted features $\phi'_{\mathcal{V}}(\mathcal{X}')$ will also be affected. Therefore, λ should be a relatively large value ranging from 0 to 1 such that the original features account for a large proportion, making the extracted transformation parameter $\mathcal{A}_\theta$ more reliable. As demonstrated in Table 1, the fused features $\lambda\phi_{\mathcal{V}}(\mathcal{X}) + (1-\lambda)\phi'_{\mathcal{V}}(\mathcal{X}')$ outperform either $\phi_{\mathcal{V}}(\mathcal{X})$ or $\phi'_{\mathcal{V}}(\mathcal{X}')$, illustrating the certain complementarity between the two features.

TABLE 2. Ablation study on network structure. Re-identification rates (%) at rank-r and mAP (%).

Method	rank-1	rank-5	rank-10	mAP
	SYSU-MM01			
Baseline	24.50	53.64	68.76	27.29
$v\text{-}\mathcal{AT}$	27.94	56.67	71.21	29.04
$t\text{-}\mathcal{AT}$	25.21	54.03	69.51	26.41
$v\text{-}\mathcal{AT} + t\text{-}\mathcal{AT}$	27.77	57.06	70.50	28.93
$v\text{-}\mathcal{AT}$+*triplet*	**29.80**	**57.43**	**72.47**	**30.45**
	RegDB			
Baseline	22.32	36.09	46.35	22.68
$v\text{-}\mathcal{AT}$	24.56	39.61	48.45	24.66
$t\text{-}\mathcal{AT}$	23.43	38.12	46.51	23.19
$v\text{-}\mathcal{AT} + t\text{-}\mathcal{AT}$	24.21	39.37	48.51	24.56
$v\text{-}\mathcal{AT}$+*triplet*	**26.60**	**40.43**	**49.15**	**26.37**

2) ABLATION STUDY ON NETWORK STRUCTURE

As demonstrated in Table 2 that the new fusion features obtained after affine transformation correction for visible images(abbreviated as v-$\mathcal{AT}$) are more discriminative, and finally rank-1, rank-5, rank-10 and mAP are significantly better than baseline on the SYSU-MM01 dataset and RegDB dataset at least 3.2% and 2.4%, respectively.

When the affine transformation correction is only performed on the infrared images(abbreviated as t-$\mathcal{AT}$), but the performance is not improved obviously. In addition, when the affine transformation correction is performed on the visible images and the infrared images at the same time(referred to as v-$\mathcal{AT}$ + t-$\mathcal{AT}$), compared with only performing the affine transformation correction on visible images(abbreviated as v-$\mathcal{AT}$), the experimental results are almost the same. We speculate the main reason is that whether the SYSU-MM01 dataset or the RegDB dataset, there are mainly some image misalignment phenomena in the visible images, whereas the infrared images are basically aligned. For dataset SYSU-MM01, there are even visible images with severe deviations likes Fig. 2(c). Therefore, we finally decide to only perform affine transformation correction on the visible images to reduce the number of overall network parameters, and integrate the triplet loss and the identity loss as the overall loss (abbreviated as v-$\mathcal{AT}$ + *triplet*).

D. COMPARISON WITH THE STATE-OF-THE-ARTS

In this subsection, we select some state-of-the-art visible infrared cross-modality person re-identification(VI-REID) methods for comparison to demonstrate the superior

TABLE 3. Comparison with other cross-modality matching methods on SYSU-MM01 data set. Re-identification rates (%) at rank-r and mAP (%).

Method	rank-1	rank-10	rank-20	mAP
	SYSU-MM01			
LOMO [5]	1.75	14.14	26.63	3.48
MLBP [40]	2.12	16.23	28.32	3.86
HOG [39]	2.76	18.25	31.91	4.24
GSM [41]	5.29	33.71	52.95	8.00
One-stream [25]	12.04	49.68	66.74	13.67
Two-stream [25]	11.65	47.99	65.50	12.85
Zero-Padding [25]	14.80	54.12	71.33	15.95
TONE [27]	12.52	50.72	68.60	14.42
HCML [27]	14.32	53.16	69.17	16.16
BCTR [26]	16.12	54.90	71.47	19.15
BDTR [26]	17.01	55.43	71.96	19.66
cmGAN [28]	26.97	67.51	80.56	27.80
D^2RL [29]	28.9	70.6	82.4	29.2
ours(MAPAN)	**29.80**	**72.47**	**84.21**	**30.45**

TABLE 4. Comparison with other cross-modality matching methods on RegDB data set. Re-identification rates (%) at rank-r and mAP (%).

Method	rank-1	rank-10	rank-20	mAP
	RegDB			
LOMO [5]	0.85	2.47	4.10	2.28
MLBP [40]	2.02	7.33	20.90	6.77
HOG [39]	13.49	33.22	43.66	10.31
GSM [41]	17.28	34.47	45.26	15.06
One-stream [25]	13.11	32.98	42.51	14.02
Two-stream [25]	12.43	30.36	40.96	13.42
Zero-Padding [25]	17.75	34.21	44.35	18.90
TONE [27]	16.87	34.03	44.10	14.92
HCML [27]	24.44	47.53	56.78	20.80
ours(MAPAN)	**26.60**	**49.15**	**61.58**	**26.37**

performance of the proposed method, including Zero-Padding [25], HCML [27],TONE [27], BCTR [26], BDTR [26], cmGAN [28] and D^2RL [29]. In addition, several other competitive methods are also included for the comparisons. The additional methods contain four feature learning-based methods: HOG [39], LOMO [5], one-stream and two-stream networks [25] and two matching model learning methods: MLAPG [40], GSM [41]. Most of the comparison results are originated from [27] on the RegDB dataset and [25] on the SYSU-MM01 dataset.

The results shown in Table 3 and Table 4 demonstrate that the proposed end-to-end learning framework(MAPAN) outperforms most existing state-of-the-art methods on the two datasets. On the SYSU-MM01 dataset, Compared to the dual-path network with a novel bi-directional dual-constrained top-ranking loss(BDTR [26]) and the two-stage feature learning and metric learning method(TONE + HCML [27]), we consistently outperform them with almost 15% for rank-1 matching rate and 10% for mAP. For the latest approaches the novel cross-modality generative adversarial network (abbreviated as cmGAN [28]) and the novel Dual-level Discrepancy Reduction Learning (D^2RL [29]) scheme, we also outperform them with at least 1% for both rank-1 matching rate and mAP. Specifically, we achieve rank-1 = 29.80% and mAP = 30.45% on the SYSU-MM01 dataset.

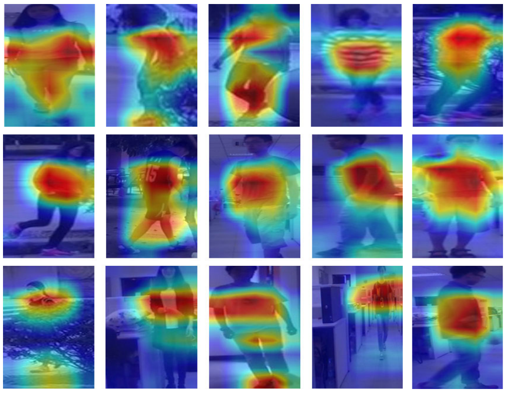


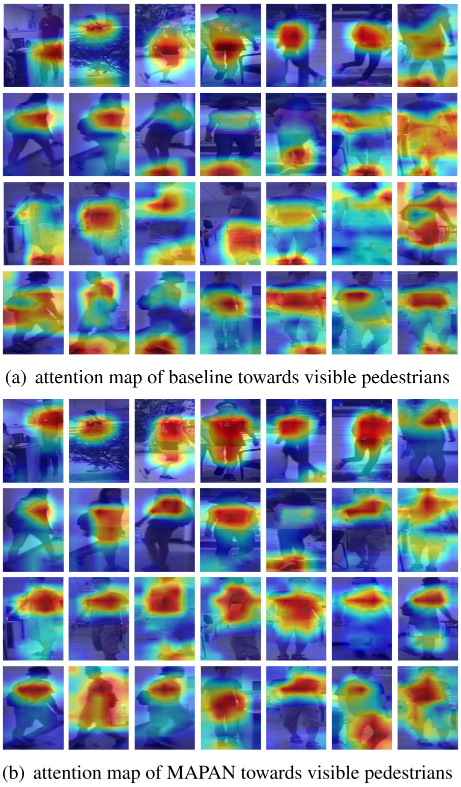
FIGURE 5. The visualization of the visible base branch's attention towards the input visible pedestrians. We can observe that the attention of the branch's high responses are mostly concentrated on the pedestrians' bodies, although the input images are more or less with occlusion, excessive background, scale variations, etc. So we can use the high responses Res5 block features to regress the affine transformation parameter $\mathcal{A}_\theta$. We achieve this visualization by Grad-CAM [37].

For the RegDB dataset, its size is smaller than the SYSU-MM01 dataset, and there is relatively less image misalignment in the RegDB dataset. Since the proposed method (MAPAN) mainly focuses on the image misalignment of large-scale datasets, it has relatively limited performance on the small-scale dataset RegDB, of which performance is not as good as those on the dataset SYSU-MM01, but is still superior to most mainstream algorithms such as TONE, HCML, etc. Specifically, we achieve rank-1 = 26.60% and mAP = 26.37% on the RegDB dataset.

E. EXPERIMENTAL RESULT VISUALIZATION

To visualize the aligned visible images, we extract the affine transformation parameter $\mathcal{A}_\theta$ and then apply the affine transformation to the visible images of SYSU-MM01 dataset. As shown in Fig.2, Fig.2(a) and Fig.2(b) are incomplete images and MAPAN tends to zoom out and rotate the images slightly. Fig.2(c), Fig.2(d) are excessive background images and Fig.2(e), Fig.2(f) are occluded images. For the two kinds of misaligned images, MAPAN tends to enlarge and translate the images, making its attention more focused on the pedestrian area. We observe that MAPAN can not perform alignment perfectly as human does, but it more or less reduces the scale variations and position deviations of the visible images and enhances the network attention towards pedestrian bodies as shown in Fig.6, so it improves the performance of VI-REID eventually.

Fig.7 shows some sample retrieval results on the dataset SYSU-MM01. The testing set contains 96 people, 3803 infrared images for query, and 301 randomly visible images are selected as the gallery set. We use the most challenging single-shot all-search mode mentioned in [25]. The first column of each row is the input query infrared image. Each row represents the retrieval result of a query. The top1-10 of the retrieval results are sorted from left to right according to similarity scores. The green and red superscripts indicate true positives and false positives, respectively. It can be observed from the figure that the pedestrians with the same identity as the query images can be correctly retrieved to some extent.


(a) attention map of baseline towards visible pedestrians

(b) attention map of MAPAN towards visible pedestrians

FIGURE 6. Attention maps of baseline and MAPAN towards pedestrians. We can observe that the attentions of the MAPAN are more focused on the pedestrians' bodies whereas the baseline is more inclined to be distracted. The proposed MAPAN not only achieves image alignment, but also enhances the network attention towards pedestrian body areas, and so improves recognition accuracy efficiently. We achieve this visualization by Grad-CAM [37].

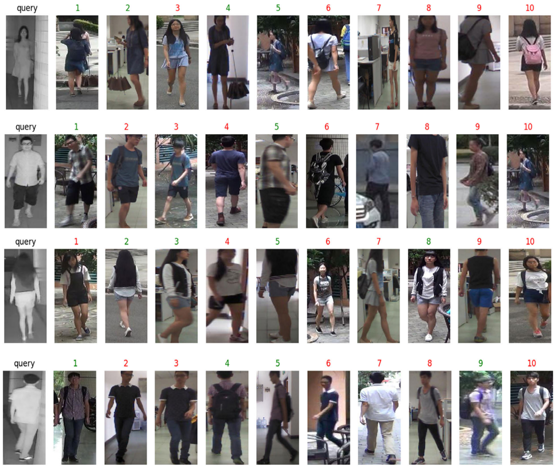


FIGURE 7. Some sample retrieval results on dataset SYSU-MM01. The first column are the input query infrared images, and the retrieved visible images are sorted according to the similarity score (cosine distance) from left to right. The green and red superscripts indicate true positives and false positives, respectively. It shows that the pedestrians with the same identity as the query images can be correctly retrieved to some extent.

V. CONCLUSION

In this paper, we propose a multi-path adaptive pedestrian alignment network (MAPAN) to learn discriminative feature representations. The multi-path network learns features directly from the input data and adaptively aligns the pedestrians without additional manual annotations. We alleviate the intra-modality discrepancies caused by image misalignment and enhance attention of the network towards pedestrians efficiently by combining the features of the adaptively aligned visible images with the features of the original visible images. To alleviate the cross-modality discrepancies between the visible domain and the infrared domain, the discriminative features of the two modalities are mapped to the same feature embedding space, and we also design a triplet loss to reduce the discrepancies, which is combined with identity loss as the overall loss. Extensive experiments illustrate that the proposed method outperforms the state-of-the-arts. In the future, we will continue to investigate the internal mechanism of adaptive pedestrian alignment network for VI-REID and consider applying more constraints to the algorithm to further improve the performance.

REFERENCES

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*. [Online]. Available: https://arxiv.org/abs/1610.02984

[2] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 152–159.

[3] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1846–1855.

[4] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1249–1258.

[5] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.

[6] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, pp. 13-1–13-20, Dec. 2017, doi: 10.1145/3159171.

[7] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch DropBlock network for person re-identification and beyond," *CoRR*, vol. abs/1811.07130, pp. 1–11, Sep. 2018. [Online]. Available: https://arxiv.org/abs/1811.07130

[8] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 667–676.

[9] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 5735–5744.

[10] S. Bai, P. Tang, P. H. S. Torr, and L. J. Latecki, "Re-ranking via metric fusion for object retrieval and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 740–749.

[11] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 393–402.

[12] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3586–3593.

[13] H. Wang, S. Gong, and T. Xiang, "Unsupervised learning of generative topic saliency for person re-identification," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Nottingham, U.K., Sep. 2014. [Online]. Available: http://www.bmva.org/bmvc/2014/papers/paper019/index.html

[14] B. Xu and G. Qiu, "Unsupervised person re-identification via graph-structured image matching," in *Proc. Comput. Vis. ACCV Workshops*, C.-S. Chen, J. Lu, and K.-K. Ma, Eds. Cham, Switzerland: Springer, 2017, pp. 301–314.

[15] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 4, pp. 83-1–83-18, Oct. 2018, doi: 10.1145/3243316.

[16] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 8738–8745.

[17] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3633–3642.

[18] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, "Unsupervised person re-identification by soft multilabel learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2148–2157.

[19] J. Meng, S. Wu, and W.-S. Zheng, "Weakly supervised person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, 2019, pp. 760–769.

[20] H. Wang, T. Fang, Y. Fan, and W. Wu, "Person re-identification based on DropEasy method," *IEEE Access*, vol. 7, pp. 97021–97031, 2019.

[21] S. Jiao, J. Wang, G. Hu, Z. Pan, L. Du, and J. Zhang, "Joint attention mechanism for person re-identification," *IEEE Access*, vol. 7, pp. 90497–90506, 2019.

[22] X. Su, X. Qu, Z. Zou, P. Zhou, W. Wei, S. Wen, and M. Hu, "k-reciprocal harmonious attention network for video-based person re-identification," *IEEE Access*, vol. 7, pp. 22457–22470, 2019.

[23] Y. Zhang, X. Gu, J. Tang, K. Cheng, and S. Tan, "Part-based attribute-aware network for person re-identification," *IEEE Access*, vol. 7, pp. 53585–53595, 2019.

[24] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.

[25] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5380–5389.

[26] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. IJCAI*, 2018, pp. 1–8.

[27] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proc. AAAI*, 2018, pp. 7501–7508.

[28] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 677–683.

[29] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 618–626.

[30] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.* 2015, pp. 2017–2025.

[31] B. Lavi, M. F. Serj, and I. Ullah, "Survey on deep learning techniques for person re-identification task," 2018, *arXiv:1807.05284*. [Online]. Available: https://arxiv.org/abs/1807.05284

[32] W. Zheng and A. Wu, "Asymmetric person re-identification: Cross-view person tracking in a large camera network," (in Chinese), *Sci. Sin Inf.*, vol. 48, no. 5, pp. 65–83, 2018.

[33] Z. Wang, Z. Wang, Y. Wu, J. Wang, and S. Satoh, "Beyond intra-modality discrepancy: A comprehensive survey of heterogeneous person re-identification," 2019, *arXiv:1905.10048*. [Online]. Available: https://arxiv.org/abs/1905.10048

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[35] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," Univ. Montréal, Montréal, QC, Canada, Tech. Rep. 1341, Jan. 2009.

[36] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

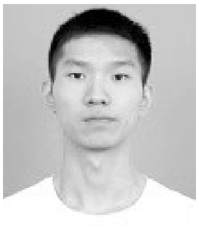
[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[38] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *CoRR*, vol. abs/1703.07737, 2017. [Online]. Available: http://arxiv.org/abs/1703.07737

[39] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[40] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *Proc. ICCV*, Dec. 2015, pp. 3685–3693.

[41] L. Lin, G. Wang, W. Zuo, X. Feng, and L. Zhang, "Cross-domain visual matching via generalized similarity measure and feature learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1089–1102, Jun. 2017.



BO LI is currently pursuing the M.S. degree with the College of Electronics and Information Engineering, Sichuan University, Chengdu, China. His main research interests include computer vision and pattern recognition.

XIAOHONG WU received the Ph.D. degree in communication and information system from Sichuan University, Chengdu, China, in 2008. She is currently an Associate Professor with the College of Electronics and Information Engineering, Sichuan University. Her research interests include image processing and pattern recognition.

QIANG LIU is currently pursuing the Ph.D. degree in signal and information processing with the College of Electronics and Information Engineering, Sichuan University. His research interests include image processing, machine learning, and computer vision.

XIAOHAI HE (M'16) received the B.S. and M.S. degrees in electrical engineering and the Ph.D. degree in biomedical engineering from Sichuan University, Chengdu, China, in 1985, 1991, and 2002, respectively. He is currently a Professor with the College of Electronics and Information Engineering, Sichuan University. His research interests include image processing, pattern recognition, computer vision, image communication, and software engineering. He is a Senior Member of the Chinese Institute of Electronics. He is an Editor of the *Journal of Information and Electronics Engineering* and the *Journal of Data Acquisition & Processing*.

FEI YANG (M'16) received the B.S. degree from Tsinghua University, the M.S. degree from Chinese Academy of Sciences, and the Ph.D. degree from Rutgers University. He is currently a Research Scientist with the TAL AI Lab. His research interests include image understanding, speech recognition, and large-scale machine learning.