IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 22, 2019, accepted December 30, 2019, date of publication January 13, 2020, date of current version January 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2966002

Exploring Cross-Modality Commonalities via Dual-Stream Multi-Branch Network for Infrared-Visible Person Re-Identification

DING CHENG, XIAOHONG LI, MEIBIN QI, XUELIANG LIU, CUIQUN CHEN, AND DAWEI NIU
School of Computer Science and Information Engineering, Hefei University of Technology, Anhui 230009, China

Corresponding author: Xiaohong Li (jsjlxh@hfut.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0804205, in part by the National Natural Science Foundation of China under Grant 61976076, and in part by the National Natural Science Foundation of China under Grant 61771180.

ABSTRACT Infrared-Visible person Re-IDentification (IV-ReID) is an emerging subject, which has important research significance for nighttime monitoring. Existing works focus on reducing cross-modality discrepancies, but the cross-modality discrepancy cannot be completely eliminated. Therefore, we concentrate on excavating cross-modality commonalities to handle the task. Since similar features between two modalities are possessed of cross-modality commonalities, our goal is to find more similar features in infrared and visible images. A novel Dual-stream Multi-layer Corresponding Fusion Network(DMCF) is proposed to explore more similar features between two modalities in this paper. It mainly contains three aspects. 1) We explore more similar features between two modalities by learning low-level features, Meanwhile, we also propose a method that the same level features between two modalities are correspondingly fused to reduce cross-modality discrepancies. 2) We adopt different Multi-granularity dividing methods for multi-layer features, so that it can improve the ability of the model to perceive feature details. 3) We separately calculate the loss for different-layer features. Therefore, we learn different weighting factors for the loss of different hierarchical features through Multi-task Learning, so that each branch can be fully optimized. Extensive experiments on two datasets demonstrate the superior performance compared to the state-of-the-arts.

INDEX TERMS Infrared-visible person Re-IDentification, cross-modality commonalities, multi-layer features, Multi-granularity features, Multi-task Learning.

I. INTRODUCTION

Person Re-Identification, referred to as Re-ID for short in this paper, which is a technique for determining whether a specific person is present in an image or video sequence by computer vision technology [3], [25]–[27], [49]. It has gained increasing attention in the research community due to its importance in various video surveillance and intelligent applications [3]. In recent years, Re-ID focuses on the visible field, which is given a query image/video of a person and search it out from a gallery set of images/videos captured by other surveillance equipment [40], [41]. However, in poor lighting conditions or at night, ordinary cameras cannot capture clear visible images of person, so we can only rely on infrared cameras to capture infrared images of person. However, searching visible(infrared) images of person from a gallery with infrared(visible) images of person is exactly a major difficulty in the field of Re-ID, which restricts the applicability of practical surveillance applications [42]–[44]. Therefore, it is particularly important to improve the technology that searching visible(infrared) images of person from a gallery with infrared(visible) images of person.

Visible images typically have high spatial resolution and considerable detail and chiaroscuro, thus, they are suitable for human visual perception. However, these images can be easily influenced by severe conditions, such as poor illumination, fog, and other effects of bad weather. Meanwhile, infrared images, which depict the thermal radiation of objects,

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

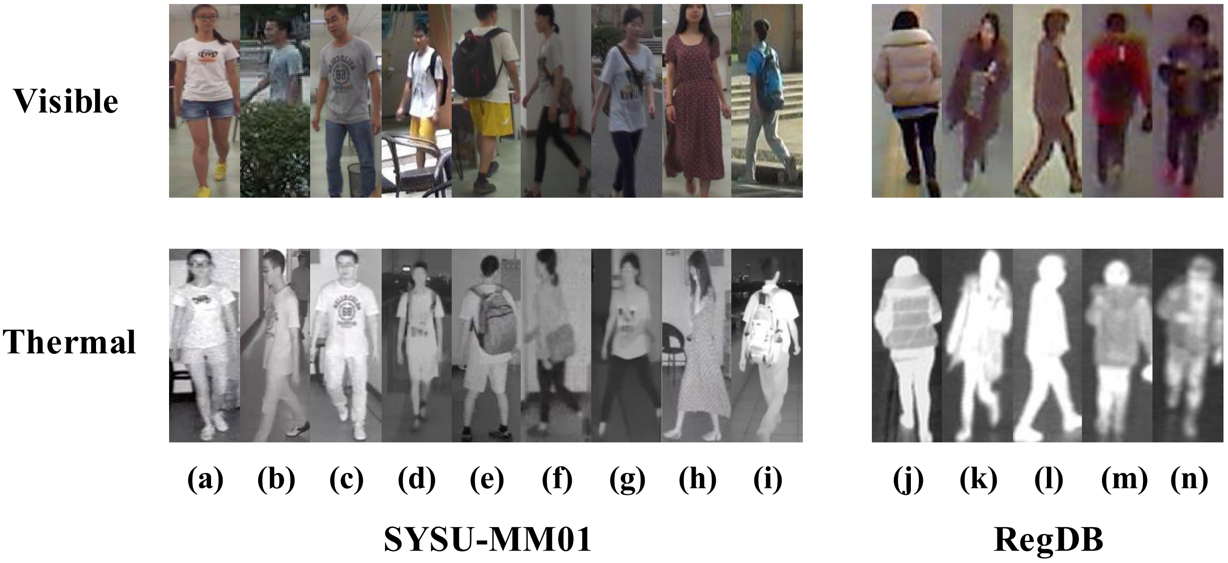


FIGURE 1. Examples of visible images and infrared images in SYSU-MM01 and RegDB datasets. Images(a-i) on the left are from SYSU-MM01 dataset and images(j-n) on the right are from RegDB dataset. Note that person images are captured by different spectrum cameras, and every columns are of the same person.

are resistant to these disturbances but typically have low resolution and poor texture [45]. If a visible (or infrared) image of a specific person is given, the system can search out the corresponding thermal (visible) images from a gallery set captured by other spectrum cameras [1]–[3]. This cross-modality image matching task is named Infrared-Visible person Re-IDentification (IV-ReID). Ye *et al.* [3] propose a Dual-path network with bi-directional Dual-constrained top-ranking loss to learn discriminative feature representations. Wang *et al.* [49] propose to reduce the modality discrepancy by unifying the image representations through image-level conversion. Firstly, these methods can reduce cross-modality discrepancies, but it is a fact that exploring cross-modality commonalities is more effective than reducing cross-modality discrepancies. Because similar features between two modalities are possessed of cross-modality commonalities, while visible and infrared images have many similarities in appearance. As shown in Fig. 1(a), both visible images and infrared images have information of obvious logo patterns. Furthermore, according to related research [46], [47], the low-level layers concentrate on appearance information to discriminate between samples, while the high-level layers contain semantic information [47]. Therefore, we believe that learning low-level features can acquire more cross-modality commonalities. Secondly, we also propose a method, the same level of features of two branches are fused, to reduce cross-modality discrepancies.

However, as shown in Fig. 1(b), visible image of person is obscured by objects, while infrared image of person is not obscured. This situation is similar to Partial Person re-identification and poses a serious challenge to the system. Therefore, we adopt the Multi-granularity dividing method which can match infrared and visible images of person through detailed information. Moreover, according to characteristic of different-layer features, we adopt different proportion dividing methods. Since low-level features contain more noise, we do not divide them, however, middle-level features and high-level features both contain large amounts of information, so we divide middle-level features into two local features and high-level features into three local features. which is more in line with the structure of the human body, such as the upper body and lower body or head, upper body and lower body. At the same time, global features of low-level features, middle-level features and high-level features are retained. Experimental results show that the dividing method is the best.

Since we extract features at different layers, our network can be regarded as a Multi-task network, we treat each different level of feature as a task. However, in Multi-task networks, some tasks tend to play a dominant role while others do not perform well. If we simply add up loss of each task, it will lead to tuning parameters of some tasks, while other task will not be fully optimized, this situation greatly affects optimization of the model. Meanwhile, researches show that performance is highly dependent on an appropriate choice of weighting between loss of each task. Searching for an optimal weighting is prohibitively expensive and difficult to resolve with manual tuning [50]. Therefore, we introduce the method of Multi-task Learning, which embeds a learnable parameter as the weight factor in the loss of each feature and balances the weight factors between loss of different-level features, so as to obtain optimal overall loss and fully optimize the network. At the same time, the method can also improve the generalization ability and robustness of the model. For example, different-layer losses may account for different proportions of overall loss in different datasets. In this case, the method can automatically adjust suitable the weight factor between loss of each task, so that the model can be fully optimized.

In a word The main contributions can be summarized as follows:

(1) We propose a method that exploring cross-modality commonalities by learning low-features to deal with cross-modality tasks.

(2) According to characteristics of different hierarchical features, we adopt different proportion dividing methods, which enables the model to capture detail features sensitively.

(3) A method of Multi-task Learning for the loss of multi-layer features is introduced. so that the model can be fully optimized.

The rest of the paper is organized as follows. Section II briefly reviews related works. Section III describes the details of our method. The experiments and discussions are shown in Section IV. We finally draw a conclusion in Section V.

II. RETATED WORK

Infrared-visible Re-ID For the IV-ReID problem, in addition to the idea of reducing the modality discrepancy, we can also deal with it by exploring the modality commonality. Existing methods attempt to reduce modality discrepancies using feature embedding frameworks similar to conventional Re-ID methods. Wu *et al.* [2] proposes a deep zero-padding framework for shared feature learning under two different modalities. Ye *et al.* [1], [22] introduce a two-step framework for feature learning and metric learning. They [3], [23], [24] also propose an end-to-end Dual-path network to learn common representations. Dai *et al.* [4] design a network to learn discriminative representations from different modalities. Wang *et al.* [49] introduce a Dual-level Discrepancy Reduction Learning (D2RL) scheme. The image-level discrepancy reduction sub-network is utilized to reduce the modality discrepancy, and the feature-level discrepancy reduction sub-network is utilized to eliminate the remaining appearance discrepancy. Although these methods can reduce cross-modality discrepancies, they ignore a fact that exploring cross-modality commonalities is more effective than reducing cross-modality discrepancies. Since more similar features which are possessed of cross-modality commonalities can be obtained by learning low-level features, we simultaneously extract low-level and middle-level features for learning when high-level features are extracted in the Dual-stream network. Meanwhile, we also propose a method, the same layer of features of two branches are fused, to reduce cross-modality discrepancies.

Multi-granularity dividing method Recently some deep Re-ID methods push the performances to a new level comparing to the former systems. Zhang *et al.* [5] introduce a part-based alignment matching in training phase with shortest path programming and mutual learning to improve metric learning performance [6], [7]. both equally slice the feature maps of input images into several stripes in vertical orientation. Reference [6] merges slices of local features with LSTM network and combine with global features learned from classification metric learning. Instead [7] directly concatenates the features from local parts as the final representation, and applies refined part pooling to modify the mapping validation of part features. In [7]–[9], images are all split into several stripes in horizontal orientation according to intrinsic human body structure knowledge, on which local feature representations are learned. [10], [11] utilize structural information of body landmarks predicted by posing estimation methods to crop more accurate region areas with semantics. To locate semantic partitions without strongly learning-based predictors, region proposal methods such as [12], [13] are employed in some part-based methods [14]–[18]. Attention information can be a powerful complement for discrimination, which are enhanced in [9], [16], [19]. In [28], the ResNet-50 backbone is split into three branches after res_conv4_1 residual block: Global Branch, Part-2 Branch and Part-3 Branch. Then, different parts are divided into horizontal stripes of different proportions as local features, and global features of each part are retained. We present a dividing method in this paper, it is optimized and ameliorated on the basis of [28]. We split different-layer features into different proportions, which is more convincing than that splitting a feature into different proportions is proposed in [28].

Multi-task learning Multi-task learning aims to improve learning efficiency and prediction accuracy for each task, when compared to training a separate model for each task [20], [21]. It can be considered an approach to inductive knowledge transfer which improves generalisation by sharing the domain information between complimentary tasks. In computer vision, Wang *et al.* [50] proposes to dynamically adjust the weight of multiple tasks, the weight of each task is balanced optimally by the homoscedastic task uncertainty. In general, Multi-task Learning is applied to different loss functions, and we propose a viewpoint that it is applied to different-layer features.

III. OUR APPROACH

In this paper, we propose a Dual-stream Multi-layer Corresponding Fusion Network(DMCF) for IV-ReID tasks. As shown in Fig. 2. It comprises two main components: 1) Dual-stream Corresponding Fusion Network, which extracts multi-layer features and fuses them. 2) Multi-granularity Multi-task loss, which splits different-layer features and obtains appropriate loss through Multi-task Learning. Specifically, low-level, middle-level, and high-level features of the Dual-stream network is extracted to obtain more similar features between two modalities and fused to reduce cross-modality discrepancies. Multi-granularity Multi-task loss ensures that detailed features can be learned and each branch of the network is fully optimized.

In this section, First of all, the importance of extracting multi-layer features and fusing the same layer of features are discussed. Secondly, the theoretical analysis of dividing features is given. Finally, we provide a detailed description of Multi-task Learning.

A. DUAL-STREAM CORRESPONDING FUSION NETWORK

Fig. 2 shows the framework of our proposed method. The backbone of our network is ResNet-50 which helps

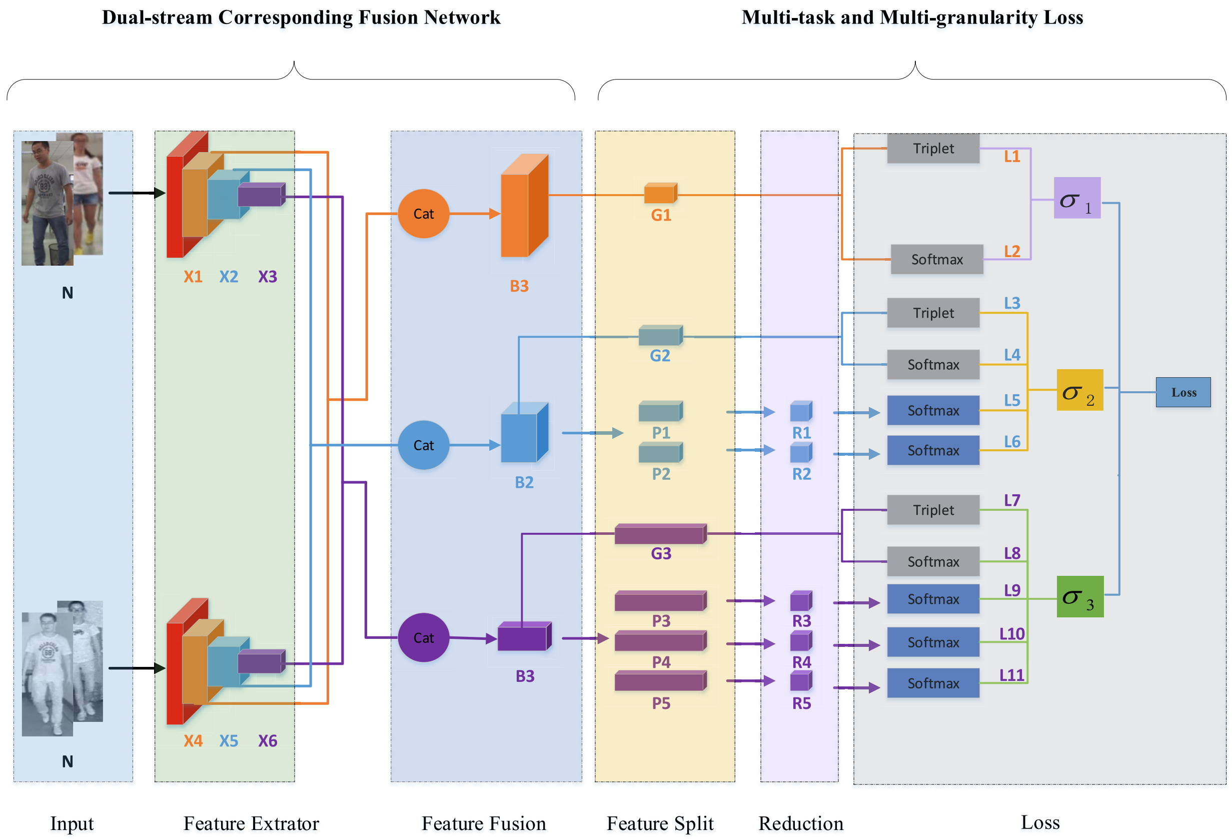


FIGURE 2. The proposed Dual-stream end-to-end learning framework for IV-ReID. "N" represents the batch size, while totally 2*N images are fed into the network for training. It comprises two main components: 1) Dual-stream Corresponding Fusion Network, which extracts multi-layer features and fuses them. 2) Multi-granularity Multi-task loss, which splits different-layer feature maps and obtains appropriate loss through Multi-task Learning. Note that "X1, X2, X3, X4, X5, X6" respectively represent different-layer features of infrared and visible images, "B1, B2, B3" respectively represent different-layer features after fusion, "G1, G2, G3" respectively represent global features after MaxPool, "P1, P2, P3, P4, P5" respectively represent local features after MaxPool and dividing, "R1, R2, R3, R4, R5" respectively represent local features after reduction, "σ_1", "σ_2" and "σ_3" respectively represent noise parameter which is the weight factor of loss of different-layer features, "L1, L2, L3, L4, L5, L6, L7, L8, L9, L10, L11" represent the loss of each branch respectively.

to achieve competitive performances in some Re-ID systems [28]. Previous method only considered reducing the cross-modality discrepancy to deal with IV-ReID tasks. However, it is a fact that exploring more cross-modality commonalities is key for IV-ReID tasks. Since similar features between two modalities are possessed of cross-modality commonalities, we aim to look for more similar features between infrared and visible images. Fristly, infrared images and visible images have a higher similarity in appearance, such as backpack, glasses, clothing patterns and so on. Secondly, later layers in CNNs are at a coarser resolution, and may not see fine-level details such as patterns on clothes, facial features, subtle pose differences etc., [31]. Thus, we simultaneously extract low-level and middle-level features for learning when high-level features are extracted in the Dual-stream network. We extract outputs of block_2, block_3, and block_4 layers as low-level features, middle-level features, and high-level features in this paper.

Next, we fuse Correspondingly low-level, middle-level, and high-level features of two branches in the batch dimension, features after fusion are represented by B1, B2 and B3 respectively. The formula is as follows.

$$B1 = concatenate((X1, X4), 0) \tag{1}$$

$$B2 = concatenate((X2, X5), 0) \tag{2}$$

$$B3 = concatenate((X3, X6), 0) \tag{3}$$

Firstly, the way we fuse is vector concatenation. and the most obvious advantage is that the data dimensions remain the same, so the information we get will not be lost. Secondly, fusion features contain two benefits: 1) It can enhance common features between infrared and visible images. 2) since unique information of infrared and visible features is retained, more information can be obtained and cross-modality discrepancies are reduced.

B. MULTI-GRANULARITY DIVIDING METHOD

First, even if no explicit attention mechanisms are imposed to enhance the preferences to some salient components, the deep network can still learn the preliminary distinction of response preferences on different body parts according to their inherent semantic meanings [28].

Second, the most significant appearance feature can be captured by global features, while the details of the image can be captured by local features. It has been widely used to divide the image features into local features and global features. There are many advantages to make use of different features comprehensively in practical application. The most important one is the complementary advantages between local features and global features. Therefore, we propose a method for dividing features.

Howeve, low-level features have higher resolution and contain more location and detail information, but they have lower semantics and more noise due to less convolution. High-level features have more information, but poor perception of details. Therefore, we don't divide B1, we divide B2 into two local features, we divide B3 into three local features. The formula is as follows.

$$B2 \longrightarrow \{P1, P2\} \tag{4}$$
$$B3 \longrightarrow \{P3, P4, P5\} \tag{5}$$

At the same time, the global feature of B1, B2 and B3 are retained. The formula is as follows.

$$\{G1, G2, G3\} = MaxPool\{B1, B2, B3\} \tag{6}$$

Finally, we reduce the dimension of local features {P1,P2,P3,P4,P5} to reduce redundant information. The formula is as follows.

$$\{R1, R2, R3, R4, R5\} = Reduction\{(P1, P2, P3, P4, P5), 256\} \tag{7}$$

C. LOSS FUNCTION

We employ both softmax loss for classification and triplet loss which is commonly used in metric learning [29], [30] and recently introduced to person Re-ID [8], [32] as loss function for global features. The formula is as follows.

$$\{L1, L3, L7\} = Triplet\_Loss\{G1, G2, G3\} \tag{8}$$
$$\{L2, L4, L8\} = Softmax\_Loss\{G1, G2, G3\} \tag{9}$$

{L1, L3, L7} respectively represent triplet loss for global features of three-layer features, {L2, L4, L8} respectively represent softmax loss for global features of three-layer features.

However, we only use softmax loss as the loss function for local features, Triples loss does not apply to local features, as loss can vary dramatically for misaligned or other problems, which causes destroy the model during training phases. The formula is as follows.

$$\{L5, L6\} = Softmax\_Loss\{R1, R2\} \tag{10}$$
$$\{L9, L10, L11\} = Softmax\_Loss\{R3, R4, R5\} \tag{11}$$

{L5, L6} respectively represent softmax loss for local features of middle-level features, {L9, L10, L11} respectively represent softmax loss for local features of high-level features.

Because learning low-level features can acquire more similar features between infrared and visible images, we extract multi-level features in the network. However, we are faced with the problem that the scale of loss of different layers varies greatly. If simple addition is adopted, the overall loss function will not be optimal. Therefore, we adopt the method of Multi-task Learning which combine multiple loss of different-layer features to simultaneously learn multiple objectives using homoscedastic uncertainty that is interpreted as task-dependent weighting [50]. Specifically, the model learns three noise parameters that are integrated into the loss of each task respectively, and three noise parameters are regarded as the weight factor of loss of low-level features, middle-level features and high-level features respectively. Then, we add up all the losses that have been properly weighted to get the optimal overall loss, so as to achieve the purpose of optimizing features at different layers. Therefore, our final overall loss function is:

$$\zeta_{B1} = L1 + L2 \tag{12}$$
$$\zeta_{B2} = L3 + L4 + L5 + L6 \tag{13}$$
$$\zeta_{B3} = L7 + L8 + L9 + L10 + L11 \tag{14}$$
$$\zeta_{total} = \frac{1}{2\sigma_1^2}\zeta_{B1} + \frac{1}{2\sigma_2^2}\zeta_{B2} + \frac{1}{2\sigma_3^2}\zeta_{B3} + \log\sigma_1 + \log\sigma_2 + \log\sigma_3 \tag{15}$$

σ is a learnable and observed noise scalar, and the value range of $\log\sigma^2$ is -2.0 to 5.0. σ_1, σ_2 and σ_3 respectively represent noise parameter which is the weight factor of loss of different-layer features, ζ_{B1}, ζ_{B2} and ζ_{B3} respectively represent loss of different-layer features, including softmax loss and triple loss.

The method of Multi-task Learning we refer to is as follows [50].

We derive a multi-task loss function based on gaussian likelihood maximization by depending on homoscedastic uncertainty. Let $f^W(x)$ be the output of a neural network with weights W on input x. We define the following probabilistic model. For the regression task, likelihood is defined as a Gaussian whose mean value is given by the model output:

$$p(y \mid f^W(x)) = N(f^W(x), \sigma^2) \tag{16}$$

with an observation noise scalar σ. For classification, we often compress the model output through a softmax function and sample from the generated probability vector:

$$p(y \mid f^W(x)) = softmax(f^W(x)) \tag{17}$$

In the case of multiple model outputs, we often define the likelihood to factorise over the outputs, given some sufficient statistics. We define $f^W(x)$ as our sufficient statistics, and obtain the following multi-task likelihood:

$$p(y_1, \cdots, y_k \mid f^W(x)) = p(y_1 \mid f^W(x)) \cdots p(y_k \mid f^W(x)) \tag{18}$$

with model outputs $y_1 \cdots y_k$.

In maximum likelihood inference, we maximise the logarithmic likelihood of the model. For example, the logarithmic likelihood in regression can be written as:

$$\log p(y \mid f^{W}(x)) \propto -\frac{1}{2\sigma^{2}}||y - f^{W}(x)||^{2} - \log \sigma \qquad (19)$$

for a Gaussian likelihood (or similarly for a Laplace likelihood) with σ observed noise parameter—capturing how much noise we have in the outputs. Then, we maximise the log likelihood of the model parameters W and observation noise parameter σ.

Let us now assume that our model output is composed of two vectors y_1 and y_2, each following a Gaussian distribution:

$$\begin{aligned} p(y_1, y_2|f^{W}(x)) &= p(y_1|f^{W}(x))\cdot p(y_2|f^{W}(x)) \\ &= N(y_1 : f^{W}(x), \sigma_1^2)\cdot N(y_2 : f^{W}(x), \sigma_2^2) \qquad (20) \end{aligned}$$

This leads to the minimisation objective, $\zeta(W, \sigma_1, \cdots, \sigma_i)$, (our loss) for our multi-output model:

$$\begin{aligned} &\zeta(W, \sigma_1, \cdots, \sigma_i) \\ &\quad = -\log p(y_1 \cdots y_i \mid f^{W}(x)) \\ &\quad = \log p(y_1|f^{W}(x)) \cdots \log p(y_i|f^{W}(x)) \\ &\quad = \log N(y_1 : f^{W}(x), \sigma_1^2) \cdots \log N(y_i : f^{W}(x), \sigma_i^2) \\ &\quad \propto -\frac{1}{2\sigma_1^2}||y_1 - f^{W}(x)||^2 - \log \sigma_1 \cdots \\ &\qquad -\frac{1}{2\sigma_i^2}||y_i - f^{W}(x)||^2 - \log \sigma_i \\ &\quad = \sum_i \frac{1}{2\sigma_i^2}\zeta_i(W) + \log \sigma_i \qquad (21) \end{aligned}$$

where we write $\zeta(W)$ for the loss of the output variable in each branch, and σ for weighting factor of the loss in each branch.

IV. EXPERIMENTS

This section reports the experiment settings, implementation details, comparisons with other methods, ablation studies and disscusions and comparisons of our method.

A. EXPERIMENT SETTINGS

1) DATASETS

We evaluate our method on two publicly available datasets: SYSU-MM01 [2] and RegDB [33].

- SYSU-MM01 [2]. It is a large-scale dataset collected by six cameras (four visible and two near-infared), including both indoor and outdoor environments. It contains in total 491 persons, and each person is captured by at least two different cameras. Following [2], we adopt the most challenging single-shot all-search mode evaluation protocol. The training set contains 395 persons, with 22,258 visible images and 11,909 infrared images, The testing set contains 96 persons, with 3,803 infrared images for query and 301 randomly selected visible images as the gallery set.
- RegDB [33]. It is collected from two aligned cameras (one visible and one far-infared). It contains totally 412 persons. Each person has 10 visible images and 10 far-infrared images, We follow the evaluation protocol in [3] to randomly split the dataset into two halves, which are used for training and testing respectively.

2) EVALUATION METRICS

The standard Cumulative Matching Characteristics (CMC) curve and mean Average Precision (mAP) are adopted to evaluate the performance. Note that there is a slight difference with the conventional Re-ID problem [34]. Images from one modality are used as the gallery set while the ones from the other modality as the probe set during testing.

B. IMPLEMENTATION DETAILS

1) NETWORK ARCHITECTURE

The architecture of our method is shown in Fig. 2. Random cropping is utilized for data argumentation, where both visible and infrared images are firstly resized to 256 × 256 × 3, and then a random cropped 288 × 144 × 3 image is fed into the network.

2) TRAINING STRATEGY

First of all, we implemente our algorithm with Pytorch and train them in an end-to-end manner. Second, channels of three different-layer features are 512, 1024 and 2048 respectively. we unify them with 1 × 1 convolution and convert them into 256 channels. Third, the pre-defined margin for the triplet loss is set as 1.8. We use Adam [33] to optimize the model and set the learning rate of feature extraction network to 0.001, set the learning rate of Multi-granularity Multi-task network to 0.01, and the momentum terms $\beta = 0.9$. The last, our training steps for the RegDB [33] dataset are 5000 and the SYSU-MM01 [2] dataset are 50000.

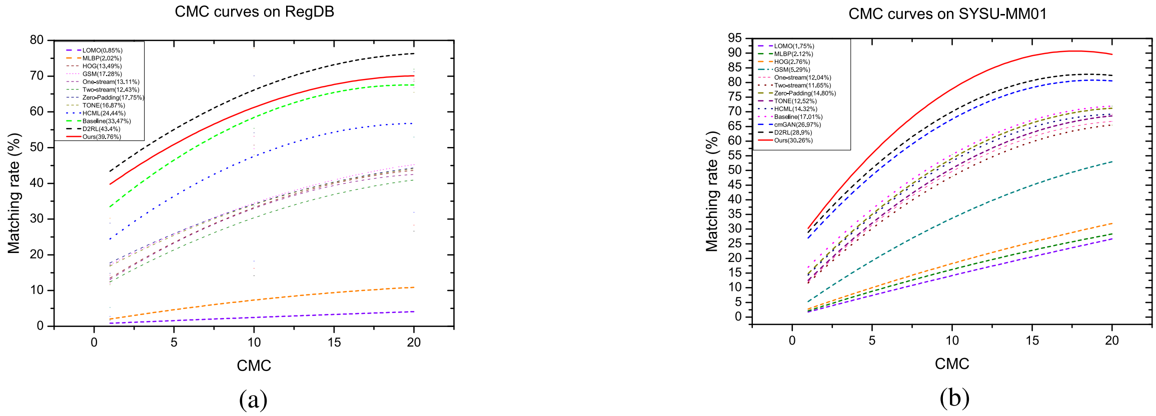
C. COMPARISON WITH THE STATE-OF-THE-ART METHODS

To demonstrate the effectiveness of our method, we compare our method with most of the related methods for IV-ReID. These methods include Zero-Padding [2], TONE [1], HCML [1], BDTR [3] and cmGAN [4]. In addition, several other learning-based methods are also included for comparisons. The additional competing methods contain some feature learning methods including HOG [35], LOMO [36], one-stream and two-stream networks [2]. The one-stream and two-stream networks are modifications of the IDE method [37] under IV-ReID settings. The detailed descriptions can be found in [2]. In addition, two matching model learning methods, MLAPG [38] and GSM [39], are also included for comparisons.

Table 1 and Fig. 3 present the results of all the methods. These methods specially designed for IV-ReID generally perform much better than the ones that are not designed for IV-ReID. First, our method significantly outperforms the state-of-the-art IV-ReID methods on the

TABLE 1. Comparison with the state-of-the arts on the RegDB and SYSU-MM01 datasets. Re-Identifification rates (%) at CMC and mAP (%).

	RegDB				SYSU-MM01			
Methods	CMC-1	CMC-10	CMC-20	mAP	CMC-1	CMC-10	CMC-20	mAP
LOMO [36]	0.85	2.47	4.10	2.28	1.75	14.14	26.83	3.48
MLBP [38]	2.02	7.33	10.90	6.77	2.12	16.23	28.32	3.86
HOG [1]	13.49	33.72	43.66	10.31	2.76	18.25	31.91	4.24
GSM [39]	17.28	34.47	45.26	15.06	5.29	33.71	52.95	8.00
One-stream [2]	13.11	32.98	42.51	14.02	12.04	49.68	66.74	13.67
Two-stream [2]	12.43	30.36	40.96	13.42	11.65	47.99	65.50	12.85
Zero-Padding [2]	17.75	34.21	44.25	18.90	14.80	54.12	71.23	15.95
TONE [1]	16.87	34.03	44.10	14.92	12.52	50.72	68.60	14.42
HCML [1]	24.44	47.53	56.78	20.80	14.32	53.16	69.17	16.16
BDTR(Baseline) [3]	33.47	58.42	67.52	31.83	17.01	55.43	71.97	19.66
cmGAN [4]	–	–	–	–	26.97	67.51	80.56	27.80
D2RL [49]	43.4	66.1	76.3	44.1	28.9	70.6	82.4	29.2
Ours	**39.75**	**61.26**	**70.10**	**40.79**	**30.26**	**75.59**	**88.13**	**33.38**



(a)

(b)

FIGURE 3. Comparison with the state-of-the arts on the RegDB and SYSU-MM01 datasets. (a) is the result on the RegDB dataset. (b) is the result on the SYSU-MM01 dataset.

TABLE 2. Results with different settings on the RegDB and SYSU-MM01 datasets.

	RegDB		SYSU-MM01	
Methods	CMC-1	mAP	CMC-1	mAP
Baseline	33.47	31.83	17.01	19.66
DCF	25.58	26.72	22.70	25.47
NO SPLIT	30.29	32.14	27.18	30.34
NO MTL	34.51	37.28	28.06	31.24
Ours	**39.75**	**40.79**	**30.26**	**33.38**

SYSU-MM01 [2] dataset. Second, the image resolution of RegDB [33] dataset is low, this is not conducive to exploring similar features between two modalities by extracting low-layer features. Meanwhile, our method is only slightly lower than D2RL method on RegDB [33] dataset. However, the net-complexity of D2RL which building GAN network framework for IV-ReID is much higher than that our method is proposed in this paper.

D. ABLATION STUDY

This subsection evaluates the proposed method with different variants, where the results on the RegDB [33] and SYSU-MM01 [2] datasets are shown in Table 2 and Fig. 4. DCF refers to extract high-level features instead of multi-layer features. NO SPLIT refers to not split features. NO MTL refers to not adopt Multi-task Learning. The Baseline in the Table 2 is BDTR [3].

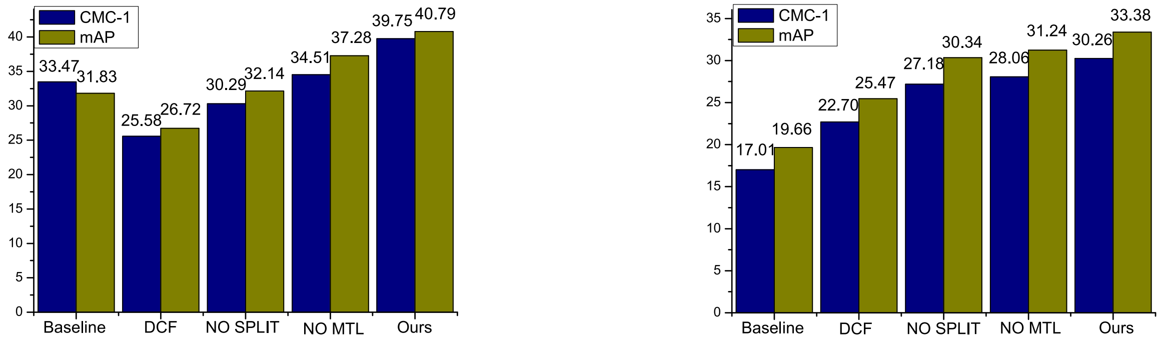
As demonstrated in Fig. 4, the CMC-1 matching rate is about 22.70% for DCF while the mAP is about 25.47% on the SYSU-MM01 [2] dataset. After introducing low-level features, our final full model could achieve CMC-1 = 30.26%, and mAP = 33.38%. The results illustrate that exploring more similar features between two modalities could improve the performance by learning low-level features.

The result shown in Table 2 illustrate that dividing features could consistently improve the performance of CMC-1 matching rate by 8-9% on the RegDB [33] dataset. It verifies the idea that dividing features helps to capture detail information which is important for identifying person.

As demonstrated in Table 2 and Fig. 4, when we introduce method of Multi-task learning, the CMC-1 matching rate is increased by 2-3% on SYSU-MM01 [2] dataset, while the CMC-1 matching rate is increased by 4-5% on RegDB [33] dataset. It strongly proves that the recognition rate of the model can be improved by fully optimizing each branch through Multi-task Learning.

E. DISCUSSIONS

Why do we extract low-level appearance features? Previous methods try to reduce cross-modality discrepancies as much as possible, but do not eliminate them completely, and we pay more attention to explore more similar features between



(a) (b)

FIGURE 4. **Evaluation of different variants of proposed method. (a) is the result on the RegDB dataset. (b) is the result on the SYSU-MM01 dataset. " DCF" refers to extract high-level features instead of multi-layer features. "NO SPLIT" refers to not split features. "NO MTL" refers to not adopt Multi-task Learning. The Baseline refers to BDTR.**

(a) (b)

FIGURE 5. **Feature visualization. "B1", "B2" and "B3" respectively represent visualization of low-level, middle-level and high-level features which are extracted in infrared and visible branch in Dual-stream ResNet50 network. "H" refers to visualization of high-level features of BDTR which are extracted in infrared and visible branch in Dual-stream AlexNet network. "Infrared" represents infrared features, "visible" represents visible features. Note that input images of (a) are from RegDB dataset and input images of (b) are from SYSU-MM01 dataset.**

two modalities. We find that infrared images and visible images have higher modal commonalities in appearance, and low-level features pay more attention to appearance information. In order to fully demonstrate the validity of low-level features, three different levels of features in our network and high-level features that are extracted in the BDTR are compared through Feature visualization. As shown in Fig. 5.

In order to improve the visual effect of features, the input image size is adjusted by image scaling in this experiment. As shown in Fig. 5, B1, B2 and B3 respectively represent three different-layer features that are extracted in this paper, H represents the high-level features that are extracted in BDTR. Based on the comparison of Feature visualization between infrared images and visible images in Fig. 5. It is not difficult to find that the similarity between VB1 and IB1 is much higher than that between VB2 and IB2, VB3 and IB3 and VH and IH. Moreover, low-level features pay more attention to the discernable information such as the glasses and the logo on the clothes in the picture. Therefore, we argue that similar features between infrared and visible images can be obtained by learning low-level features. so as to improve model recognition rate.Through feature visualization, it strongly proves the necessity of extracting low-level features.

Why do we adopt three levels of multi-granularity features. First of all, the lowest level features contain more common information, but the noise is also more, if features of layer_1 are introduced, the model will be destroyed, so the four-level features cannot be adopted. Secondly, if two levels of features are adopted, their common features cannot be fully explored, so we cannot adopt features of two levels. As shown in the Table 3, the comparison experiment fully proves our conjecture. At the same time, the experiment shows that middle-level features perform better, because middle-level features have less noise and contain abundant common features.

TABLE 3. **Results with Extracting different multi-level features on the RegDB and SYSU-MM01 datasets.**

	RegDB		SYSU-MM01	
Methods	CMC-1	mAP	CMC-1	mAP
Baseline	33.47	31.83	17.01	19.66
feature_1_2_3_4	23.76	25.69	20.89	22.66
feature_2_4	27.91	27.38	22.69	25.98
feature_2_3	36.88	38.64	28.36	31.56
feature_3_4	37.89	39.54	28.98	31.54
feature_2_3_4	**39.75**	**40.79**	**30.26**	**33.38**

Through experiments, we find that if Multi-task Learning is introduced at the beginning, some loss will be given a relatively small weight factor, so that the convergence rate of model will be slower. Thus, after the model converges, Multi-task Learning was introduced. The dynamic changes of its parameters are shown in Fig. 6. It can be seen from the Fig. 6

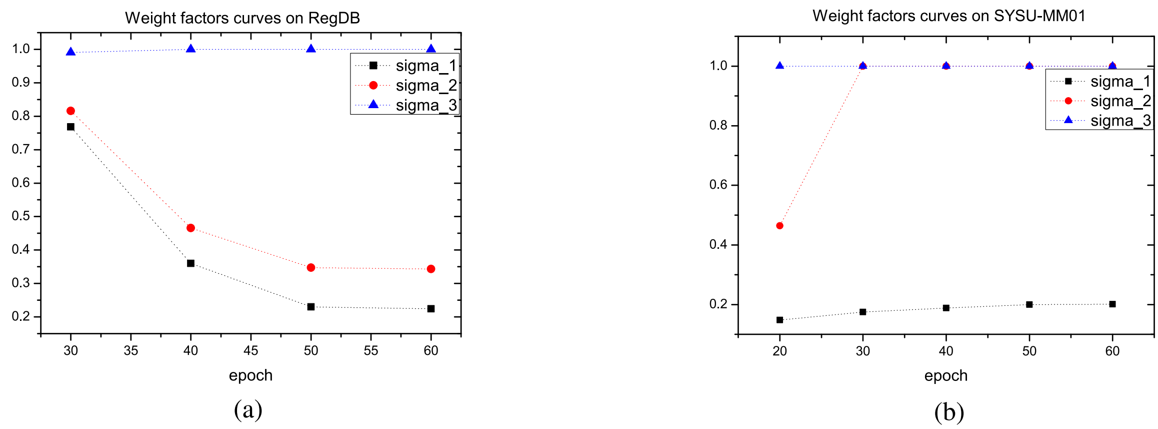


FIGURE 6. Weight factor Curves. (a) is the Curves on the RegDB dataset. (b) is the Curves on the SYSU-MM01 dataset.

that the ratio of final three noise parameters is 0.2:1:1 on the SYSU-MM01 dataset and the ratio of final three noise parameters is 0.2:0.3:1 on the RegDB dataset.

V. CONCLUSION

In this paper, we propose a Dual-stream Multi-layer Corresponding Fusion Network(DMCF) for IV-ReID tasks. In order to handle cross-modality variations, different-layer features of the two branches are extracted in the Dual-stream network, so that more similar features between infrared and visible images are acquired. Meanwhile, we propose a way of fusing features at the same layer to reduce cross-modality discrepancies and enhance the discriminability of learnt representations. At the same time, in order to imporve the sensitivity to capture details of the image, we propose a Multi-granularity dividing method that feature is splited into different scales for different-layer features. In addition, we introduce Multi-task Learning to give reasonable relative weights between the loss of different-layer features, so as to obtain the optimal overall loss, so that each branch of the network is fully optimized. Extensive experiments illustrate that our method shows significant improvement against the state-of-the-art methods.

REFERENCES

[1] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identiffication," in *Proc. AAAI Conf. Artif. Intell.*, 2018.

[2] 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.

[3] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018.

[4] 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.*, Jul. 2018.

[5] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "AlignedReID: Surpassing human-level performance in person re-identification," 2017, *arXiv:1711.08184*. [Online]. Available: https://arxiv.org/abs/1711.08184

[6] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, "Deep-person: Learning discriminative deep features for person re-identification," 2017, *arXiv:1711.10658*. [Online]. Available: https://arxiv.org/abs/1711.10658

[7] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *Proc. Eur. Conf. Comput. Vis.*, 2018.

[8] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016.

[9] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017.

[10] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017.

[11] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013.

[12] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015.

[13] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Conf. Neural Inf. Process. Syst.*, 2015.

[14] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017.

[15] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.

[16] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "HydraPlus-net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017.

[17] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian, "Deep representation learning with part loss for person re-identification," 2017, *arXiv:1707.00798*. [Online]. Available: https://arxiv.org/abs/1707.00798

[18] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017.

[19] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[20] S. Thrun, "Is learning the n-th thing any easier than learning the first?" in *Proc. Adv. Neural Inf. Process. Syst.*, 1996.

[21] J. Baxter, "A model of inductive bias learning," *J. Artif. Intell. Res.*, vol. 12, pp. 149–198, Jul. 2018.

[22] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 407–419, 2020.

[23] M. Ye, Y. Cheng, X. Lan, and H. Zhu, "Improving night-time pedestrian retrieval with distribution alignment and contextual distance," *IEEE Trans. Ind. Informat..*, vol. 16, no. 1, pp. 615–624, Jan. 2020.

[24] M. Ye, X. Lan, and Q. Leng, "Modality-aware collaborative learning for visible thermal person re-identification," in *Proc. 27th ACM Int. Conf. Multimedia (MM)*, 2019.

[25] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2976–2990, Jun. 2019.

[26] X. Gu, B. Ma, H. Chang, S. Shan, and X. Chen, "Temporal knowledge propagation for image-to-video person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9647–9656.

[27] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9317–9326.

[28] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. ACM Multimedia Conf. (MM)*, 2018.

[29] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015.

[30] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016.

[31] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger, "Resource aware person re-identification across multiple resolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.

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

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

[34] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015.

[35] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2005.

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

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

[38] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015.

[39] 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.

[40] Z. Wang, R. Hu, C. Chen, Y. Yu, J. Jiang, C. Liang, and S. Satoh, "Person reidentification via discrepancy matrix and matrix metric," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 3006–3020, Oct. 2018.

[41] X. Zhu, B. Wu, D. Huang, and W.-S. Zheng, "Fast open-world person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2286–2300, Aug. 2018.

[42] X. Lan, A. J. Ma, P. C. Yuen, and R. Chellappa, "Joint sparse representation and robust feature-level fusion for multi-cue visual tracking," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5826–5841, Dec. 2015.

[43] X. Lan, S. Zhang, P. C. Yuen, and R. Chellappa, "Learning common and feature-specific patterns: A novel multiple-sparse-representation-based tracker," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2022–2037, Apr. 2018.

[44] J. Jiang, C. Chen, J. Ma, Z. Wang, Z. Wang, and R. Hu, "SRLSP: A face image super-resolution algorithm using smooth regression with local structure prior," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 27–40, Jan. 2017.

[45] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.

[46] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2014.

[47] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015.

[48] Y. Ro, J. Choi, D. Jo, B. Heo, J. Lim, and J. Choi, "Backbone can not be trained at once: Rolling back to pre-trained network for person re-identification," 2019, *arXiv:1901.06140*. [Online]. Available: https://arxiv.org/abs/1901.06140

[49] 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.*, 2019, pp. 618–626.

[50] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.

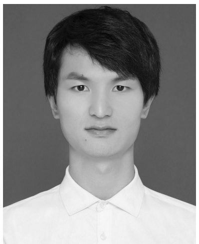
DING CHENG received the B.E. degree from Hefei Normal University. He is currently pursuing the M.E. degree with the Hefei University of Technology. His research interests include digital image analysis and processing, computer vision, and deep learning.

XIAOHONG LI is currently an Associate Professor with the School of Computer and Information, Hefei University of Technology, China. Her research interests include pattern recognition and multimedia information processing.

MEIBIN QI received the B.E. degree in radio technology from Chongqing University, in 1991, and the M.E. and Ph.D. degrees in signal and information processing from the Hefei University of Technology, China, in 2001 and 2007, respectively. He is currently a Professor with the School of Computer and Information, Hefei University of Technology. His research interests include pattern recognition, video coding, video surveillance, and the application of DSP technology.

XUELIANG LIU received the M.Sc. degree from USTC China, in 2008, and the Ph.D. degree from EURECOM France, in 2013. He is currently an Associate Professor with the School of Computer and Information, Hefei University of Technology, China. He has authored over 30 journal and conference papers in these areas, such as the IEEE TCB, ACM TOMM, ACM MM, and ICMR. His research interests include social media analysis, multimedia retrieval, and object detection. He has been serving the technical program committees of numerous multimedia and information retrieval conferences, including ACM Multimedia (MM), ACM SIGIR, the International Conference on Multimedia Retrieval (ICMR), and the International Conference on Multimedia and Expo (ICME). In addition, he is also the Co-Founder of ACM workshop on MAHCI, and served as the organizing co-chairs of ICIMCS 2013, MMM 2016, and PCM 2018.

CUIQUN CHEN received the B.E. degree from Fuyang Normal University, where she is currently pursuing the joint master's and Ph.D. degrees. Her research interests include digital image analysis and processing, computer vision, and machine learning.



DAWEI NIU received the B.E. degree from the Hunan University of Science and Technology. He is currently pursuing the M.S. degree with the Hefei University of Technology. His research interests include digital image analysis and processing and computer vision.