IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON MISSION CRITICAL SENSORS AND SENSOR NETWORKS (MC-SSN)

Received April 30, 2019, accepted May 15, 2019, date of publication May 20, 2019, date of current version May 31, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2917939

Coarse-Fine Convolutional Neural Network for Person Re-Identification in Camera Sensor Networks

ZHONG ZHANG, (Member, IEEE), HAIJIA ZHANG, AND SHUANG LIU, (Member, IEEE)
Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China

Corresponding author: Zhong Zhang (zhong.zhang8848@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61501327 and Grant 61711530240, in part by the Natural Science Foundation of Tianjin under Grant 17JCZDJC30600, in part by the Fund of Tianjin Normal University under Grant 135202RC1703, in part by the Open Projects Program of the National Laboratory of Pattern Recognition under Grant 201800002, and in part by the Tianjin Higher Education Creative Team Funds Program.

ABSTRACT In this paper, we present a novel deep model named coarse-fine convolutional neural network (CFCNN) for person re-identification in camera sensor networks, which jointly learns global and multi-scale local features simultaneously. To this end, we design the CFCNN as a multi-branch network, which is composed of one coarse and two fine branches. Specifically, the global feature is learned from the coarse branch, and the two fine branches are developed to extract two kinds of local features with different scales. Afterward, each branch is followed by a classification loss to make the identity prediction. Finally, we obtain completed pedestrian representations via concatenating the learned global and all local features. We conduct a number of experiments to evaluate the effectiveness of the CFCNN on three datasets. The CFCNN achieves high rank-1 and mAP accuracy with 94.0%/81.2%, 64.6%/58.4%, and 85.7%/72.4% on Market-1501, CUHK03, and DukeMTMC-reID, respectively. These results significantly outperform the prior state-of-the-art methods.

INDEX TERMS Person re-identification, convolutional neural network, camera sensor networks.

I. INTRODUCTION

Given a probe, person re-identification (Re-ID) aims to spot the specific person in camera sensor networks. It has been widely applied in several subfields of video security monitoring system, such as multi-camera activity analysis [1], cross-camera tracking [2] and so on. Person Re-ID is a challenging issue because of complex pedestrian images caused by variances in posture, viewpoint, illumination, background, etc.

The traditional methods for person Re-ID usually employ hand-crafted features, such as color, edge and shape, to describe the appearance of pedestrian [3]–[7]. Recently, with the prosperity of deep learning, many approaches [8]–[16] employ convolutional neural network (CNN) and obtain the breakthrough in cumulative match characteristic (CMC) curve and mean average precision (mAP). Several CNN-based methods [17]–[19] extract global features from entire pedestrian images which could represent the macroscopic clues of pedestrian. However, these methods ignore detail information of pedestrian, which is very important to discriminate different identities. The two pedestrian images in Figure 1 are with different identities, while the general appearances of them are quite similar. Hence, the global feature is difficult to distinguish them due to neglecting the discriminative region, e.g., head region.

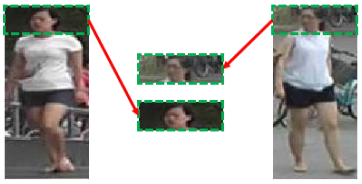

FIGURE 1. The two pedestrians with different identities have quite similar appearance in clothing. It is difficult to distinguish them only using the global feature because of neglecting the detail information in the head region.

Different from global feature learning, many researchers extract local features using different strategies in order to obtain the detail information of pedestrian. Some

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

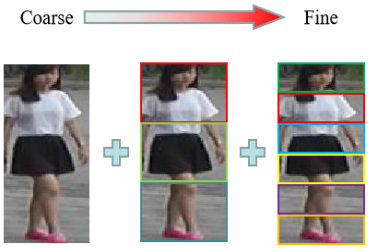

FIGURE 2. From left to right, the detail information of pedestrian increases with more local parts being partitioned. Left: the entire pedestrian image contains the global information of pedestrian appearance. Middle: the pedestrian image is partitioned into three local parts from which detail information can be extracted. Right: more local parts reflect more detail information. Completed pedestrian information exists in combination of different scales.

approaches [20], [21] directly partition pedestrian images into several fixed regions, and then learn the local feature from each region. In addition, several methods [22], [23] employ external clues, e.g., human pose estimation [24], to discover meaningful local regions, but these methods require extra supervision which is prone to error accumulation.

In this paper, a CNN-based model named Coarse-Fine Convolutional Neural Network (CFCNN) is proposed for person Re-ID in camera sensor networks, which jointly learns global and multi-scale local features simultaneously. To this end, we design CFCNN as a multi-branch network which is composed of coarse and fine branches. Specifically, the global feature is learned from the coarse branch, and the fine branch is developed to extract the local feature. From Figure 2, we can see that the detail information of pedestrian exists in different scales, and therefore we develop several fine branches to mine completed local features from various scales. In the process of learning completed local features, we directly divide convolutional activation maps into different scales, which is simple and efficient. Afterwards, each branch is followed by a classification loss for the enhancement of feature discriminative ability.

We make the following three contributions in this paper.

- Firstly, a deep model is proposed to fuse global and multi-scale local features.
- Secondly, we illuminate that learning multi-scale local features is beneficial to the enhancement of discriminative ability for pedestrian representation.
- Thirdly, experimental results on three person Re-ID datasets, i.e., Market-1501 [25], CUHK03 [26] and DukeMTMC-reID [27], show the effectiveness of CFCNN.

II. RELATED WORK

With the development of CNNs, it has been broadly used in many subfields of image classification [28], [29], such as person Re-ID and vehicle re-identification [30]. Person Re-ID has experienced a revolution from hand-crafted features to deep features. Hence, we introduce both of them in this section.

A. HAND-CRAFTED FEATURES

There was a period of prosperity for hand-crafted features [3]–[5] before deep learning methods became mainstream in person Re-ID. Gheissari *et al.* [3] utilized normalized color and salient edge histograms to represent pedestrian images, which is robust to the changes of environment and appearance. To extract discriminative features, Ma *et al.* [31] converted each pixel of a pedestrian image into a 7-dim vector containing information of coordinates, intensity, the 1st and 2nd order derivative of pixel, and then encoded them using the Fisher Vector. Hu *et al.* [32] proposed three kinds of local features, i.e., Hierarchical Weighted Histograms (HWH), Gabor Ternary Pattern HSV (GTP-HSV) and Maximally Stable Color Regions (MSCR), for pedestrian representation, which could impose the structural constraints on part-level, pixel-level and blob-level, respectively. Bazzani *et al.* [33] presented the Symmetry-Driven Accumulation of Local Features (SDALFs) to extract three complementary features for pedestrian images based on the symmetry and asymmetry perceptual principles. Yang *et al.* [34] designed the Salient Color Names based Color Descriptor (SCNCD) to compute color names distributions over different color models for addressing the illumination changes. The Local Maximal Occurrence (LOMO) [6] algorithm integrating HSV color histograms and SILTP descriptors aims to overcome the viewpoint variances by maximizing the occurrence of feature vectors of horizontal regions.

B. DEEP FEATURES

Recently, deep features [35]–[37] outperform hand-crafted features in CMC and mAP, and therefore they play a dominant role in the person Re-ID community. We catalog the deep features into three types: (1) global features, (2) local features and (3) fusion of them.

Many approaches are proposed to learn global features using entire pedestrian images. For example, Ding *et al.* [17] fed the triplet entire images into CNN, and maximized the relative distance to learn discriminative global features. In addition, Zheng *et al.* [18] presented the Pedestrian Alignment Network (PAN) to address misalignment problem by finding the optimal affine transformation when learning global features. To enhance the global feature discrimination, Geng *et al.* [19] combined the contrastive loss and the identification loss for the CNN model optimization.

Compared with the global feature, the local feature pays more attention to the detail information of pedestrian. Varior *et al.* [10] presented the Matching Gate function to compare local features along the horizontal stripe so as to learn the finer detail information of pedestrian. Ustinova *et al.* [38] learned local features by training each part of pedestrian image in a multi-region bilinear subnetwork. In order to overcome the pose variation and the background noise, Zheng *et al.* [23] introduced three types of PoseBoxes for local feature extractions from different body

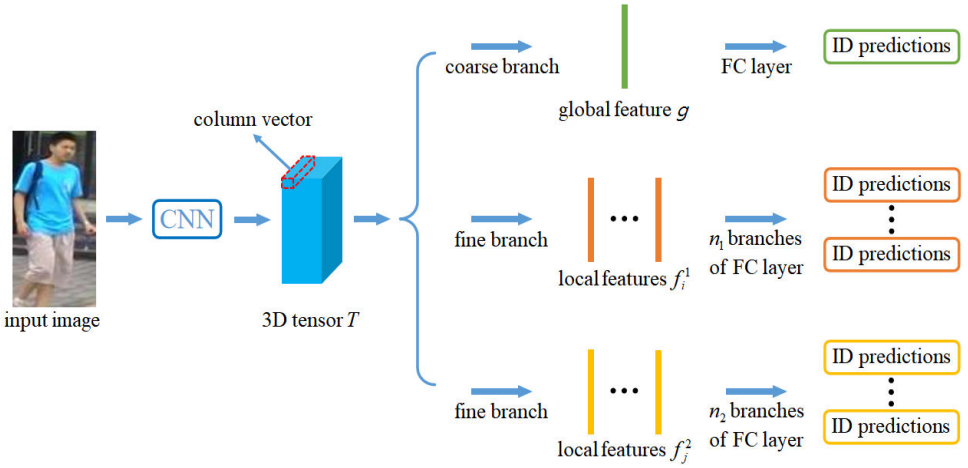

FIGURE 3. The structure of CFCNN. The input pedestrian image is firstly processed by the modified ResNet-50 to produce a 3D tensor T. The coarse branch takes the tensor T as input and outputs the global feature g. As for local features f_i^1 and f_j^2, they are extracted from the fine branches when T is divided into n_1 and n_2 uniform horizontal stripes respectively. Afterwards, g, f_i^1 and f_j^2 are respectively fed into the independent classifier and the loss function.

regions, such as arms, legs and so on. Sun *et al.* [39] directly split the feature maps of CNN into several fixed stripes, and then learned local features from them.

Some researchers fuse global and local features to take advantage of their strengths. Cheng *et al.* [40] fused global full body and local body region features via the improved triplet loss in the parts-based CNN model which contains multiple channels. Li *et al.* [41] exploited detail information of full body and local body regions in the Multi-Scale Context Aware Network (MSCAN) to mine discriminative features, and meanwhile they utilized the Spatial Transformer Networks (STN) for spatial constraints on local part-based features. Zhao *et al.* [42] captured the macro and micro-body features using Spindle Net, and merged them using the Feature Fusion Network (FFN). Zhang and Si [43] employed the verification model and the identification model for local part-based and global body-based feature extractions, and combined them using a weighted strategy.

III. APPROACH

In this section, we detail the architecture of CFCNN and the way to fuse global and multi-scale local features.

A. THE STRUCTURE OF CFCNN

1) BACKBONE NETWORK

CFCNN can take any CNN-based network as backbone, e.g., VGG [44] and ResNet [45]. Since the ResNet-50 [45] has excellent characteristics as well as the relatively succinct structure, we utilize it as backbone.

2) FROM BACKBONE TO CFCNN

We make some slight modifications on the ResNet-50. Specifically, the global average pooling and subsequent layers are removed, and meanwhile the stride of Conv5_1 layer is reset to 1 for higher spatial size. Except for the above-mentioned modifications, the rest parts of ResNet-50 remain the same as shown in Table 1. Hence, when the pedestrian image is resized to 480×160, the output is a 3D tensor T with the size of $2048 \times 30 \times 10$. Here, there are 2048 convolutional activation maps in T, and each of them is with the size of 30×10.

In order to jointly learn global and multi-scale local features simultaneously, CFCNN is designed as a multi-branch network including one coarse branch and two fine branches as shown in Figure 3. Note that we define the vector composed by activation values along the channel axis as a column vector. We utilize the coarse branch to learn the global feature. Specifically, we directly apply a global average pooling layer on T, and then employ a convolutional layer containing 256 kernels with the size of 1×1 to reduce the dimension from 2048 to 256. Hence, we obtain the global feature $g \in \mathbb{R}^{256\times1}$. Meanwhile, we employ the fine branches to learn local features. Concretely, CFCNN divides T into n_1 and n_2 uniform horizontal stripes respectively, and produces a single column vector using a local horizontal average pooling layer to average all column vectors in the same stripe. Afterwards, for each stripe we apply a convolutional layer including 256 kernels with the size of 1×1 for the dimension reduction to 256. Hence, we obtain two types of local features with different scales. We denote the local features from the fine branch with n_1 stripes as $f_i^1 \in \mathbb{R}^{256\times1}(i = 1, 2, \cdots, n_1)$. Similarly, the local features from another fine branch are denoted as $f_j^2 \in \mathbb{R}^{256\times1}(j = 1, 2, \cdots, n_2)$. It should be noticed that in order to learn different scale local features, we set n_1 and n_2 with different values. Finally, for each local feature or global feature, we make the pedestrian identity prediction by applying a FC layer followed by the softmax function independently.

TABLE 1. The architecture of the modified ResNet-50.

Layer Name	Output Size	Kernel Size	Stride	Padding
Conv1	$64 \times 240 \times 80$	$[7 \times 7, 64]$	2	(3, 3)
Max Pooling	$64 \times 120 \times 40$	3×3	2	(1, 1)
Conv2_x	$256 \times 120 \times 40$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \times 3$	$\begin{bmatrix} (0,0) \\ (1,1) \\ (0,0) \end{bmatrix} \times 3$
Conv3_x	$512 \times 60 \times 20$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \times 3$	$\begin{bmatrix} (0,0) \\ (1,1) \\ (0,0) \end{bmatrix} \times 4$
Conv4_x	$1024 \times 30 \times 10$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \times 5$	$\begin{bmatrix} (0,0) \\ (1,1) \\ (0,0) \end{bmatrix} \times 6$
Conv5_x	$2048 \times 30 \times 10$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \times 3$	$\begin{bmatrix} (0,0) \\ (1,1) \\ (0,0) \end{bmatrix} \times 3$

B. LOSS FUNCTION

In the training phase, the global feature and local features are respectively fed into the independent classifier and the loss function. Hence, they do not share the weights. Specifically, each classifier is designed to contain a FC layer followed by the softmax function, and we utilize the cross-entropy loss as the loss function in all branches. The loss function in the coarse branch is defined as:

$$L_g = -\sum_{c=1}^{C} p_c(g) \log q_c(g) \tag{1}$$

where C denotes the number of pedestrian identities, $q_c(g) \in [0, 1]$ indicates the identity prediction values that the global feature g belongs to the c-th identity, and $p_c(g)$ is the target label value of g. When g belongs to the s-th identity, then $p_s(g) = 1$; otherwise $p_c(g) = 0$. We utilize the softmax function to compute the prediction probability $q_c(g)$:

$$q_c(g) = \frac{e^{a_c}}{\sum_{m=1}^{C} e^{a_m}} \tag{2}$$

where a_c denotes the activation value of the c-th neuron in the FC layer.

The total loss function in the fine branch with n_1 stripes is formulated as:

$$L_1 = -\sum_{i=1}^{n_1} \sum_{c=1}^{C} p_c(f_i^1) \log q_c(f_i^1) \tag{3}$$

where $q_c(f_i^1) \in [0, 1]$ indicates the prediction values that f_i^1 belongs to the c-th identity, and $p_c(f_i^1)$ is the target identity label of f_i^1. Note that f_i^1 is the local feature, and it is assigned to the same label with the global feature. If f_i^1 belongs to the s-th identity, then $p_s(f_i^1) = 1$; otherwise $p_c(f_i^1) = 0$. Similar to f_i^1, the total loss of the fine branch with n_2 stripes is denoted as L_2.

In a word, the loss function of CFCNN is formulated as:

$$Loss = L_g + \lambda L_1 + \mu L_2 \tag{4}$$

where λ and μ are the coefficients to control the weight of local features. In the process of optimization, CFCNN could consider the global feature and the local features with different scales simultaneously. We update CFCNN parameters with the back-propagation and the stochastic gradient descent (SGD) algorithm.

C. FEATURE FUSION

The global and multi-scale local features extracted from CFCNN are fused to characterize the pedestrian image as shown in Figure 4. In the test stage, CFCNN produces completed descriptor by concatenating the global and all local features:

$$D = [g, f_1^1, f_2^1, \cdots, f_{n_1}^1, f_1^2, f_2^2, \cdots, f_{n_2}^2] \tag{5}$$

D. HYPERPARAMETERS

The proposed CFCNN has several key parameters, for example, the number of uniform horizontal stripes, the number of convolutional kernels, input pedestrian image size, and the coefficients of CFCNN loss. In the experiments, the hyper-parameters of CFCNN are set as follows.

(1) To learn local features, we divide the tensor T into 3 and 6 uniform horizontal stripes for two fine branches, i.e., $n_1 = 3$ and $n_2 = 6$.

(2) The number of kernels in the convolutional layer for the coarse and fine branches is set to 256 for dimension reduction.

(3) The pedestrian image is resized into 480×160 with the aspect ratio of 3 : 1.

(4) The coefficients λ and μ in Eq. 4 are both assigned to the value of 1.

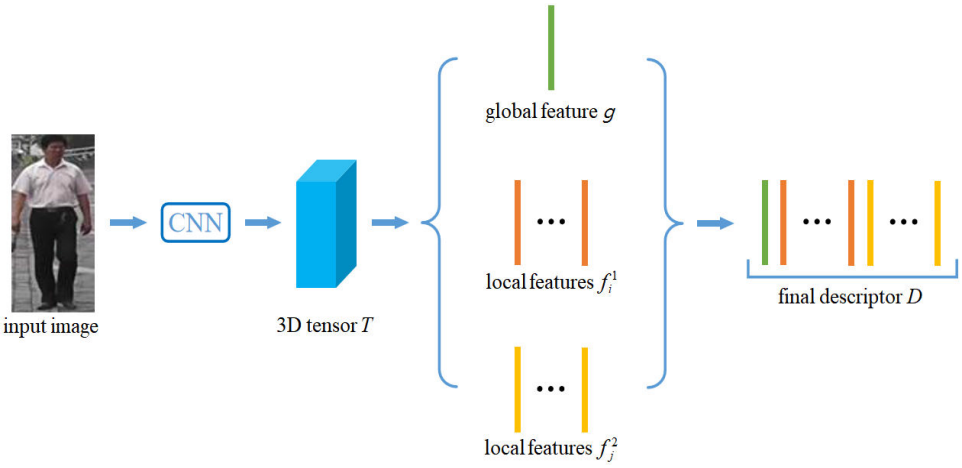

FIGURE 4. The feature representation of pedestrian image using CFCNN. The global feature g and all local features, i.e., f_i^1 and f_j^2, are concatenated to form the final descriptor D.

IV. EXPERIMENTS

This section mainly includes four aspects. We first introduce Market-1501, CUHK03 and DukeMTMC-reID datasets, and then detail the experiment settings. Afterwards, the results are reported. Finally, several key hyperparameters are analyzed in Section IV-D.

A. DATASETS

Market-1501 is composed of 32,668 images of 1,501 identities which are captured in the Tsinghua University. The pedestrian images are observed by at most six camera views. Hence there are multiple images for the same identity under different cameras. Each identity has about an average of 17 images. Concretely, the training set contains 12,936 images of 751 identities, and the test set includes 19,732 pedestrian images of 750 identities where the query set consists of 3,368 pedestrian images. Instead of manual labeled operation, the Deformable Part Model (DPM) [46] is adopted to detect all pedestrian images with consideration of the acceptable misalignment error.

CUHK03 contains 14,097 pedestrian images of 1,467 identities. Each identity is captured by two different cameras and includes about 5 pedestrian images in average for each camera. CUHK03 is divided into two parts, that is, one is the training set including 767 identities and the other is the test set containing 700 identities. The manually labeled operation and automatically labeled operation by DPM are adopted in this dataset. We evaluate CFCNN under the setting of automatically labeled operation by DPM.

DukeMTMC-reID consists of 36,411 pedestrian images from 8 high-resolution cameras. Concretely, there are 1,404 identities captured by at least 3 cameras and 408 identities (distracter identity) captured by only one camera. The training set is composed of 702 randomly selected identities, and the test set contains the rest 1,110 identities (702 identities and 408 distracter identities). In the query set, each identity under each camera view has 1 pedestrian image, and the remaining images are treated as the gallery. As a result, we obtain 16,522 training images, 2,228 query images and 17,661 gallery images.

FIGURE 5. Some pedestrian images from (a) Market-1501, (b) CUHK03, and (c) DukeMTMC-reID.

These datasets are challenging for CFCNN evaluation due to complex pedestrian images caused by variances in posture, viewpoint, illumination, background and so on. Figure 5 shows some pedestrian images from them, and Table 2 lists the statistical information. In order to evaluate CFCNN, we report the CMC at rank-1 and mAP on these datasets.

B. EXPERIMENT SETTINGS

CFCNN takes the pre-trained ResNet-50 as backbone. For each pedestrian image, we do not apply external data argumentation except random horizontal flip and normalization.

TABLE 2. The statistical information of three datasets. A/B indicates identities/images.

datasets	training set	query set	gallery set
Market-1501	751/12,936	750/3,368	750/15,913
CUHK03	767/7,365	700/1,400	700/5,332
DukeMTMC-reID	702/16,522	702/2,228	702/17,661

Concretely, the normalization is implemented by subtracting the mean values 0.486, 0.459, 0.408 and then dividing by the variance values 0.229, 0.224, 0.225 in RGB channels, respectively. In the training phase, we set the epoch number and the batch size to 80 and 64, respectively. As for the basic learning rate, we keep it at 0.1 before 40 epochs and then decayed to 0.01. Meanwhile, we set the learning rate of the pre-trained layers to 0.1 times as large as the basic learning rate during the whole training stage. The weight decay and momentum are set to 0.0005 and 0.9, respectively. It should be noticed that we adopt the same parameters on three datasets.

C. PERFORMANCE EVALUATION

We make comparison with the prior art on three datasets. Experimental results in detail are given as follows.

1) MARKET-1501

The detailed results on Market-1501 are summarized in Table 3. Concretely, the compared methods are categorized into three types, i.e., global deep features, local deep features and fusion of them. From Table 3, we can see that CFCNN obtains 94.0% and 81.2% in rank-1 and mAP accuracy respectively which exceed the other methods by a large margin. CFCNN surpasses the first type methods because it considers the structure information of pedestrian, and also outperforms the second type methods due to learning the global features. Specifically, CFCNN obtains +1.7% in rank-1 accuracy and +3.8% in mAP accuracy improvements on PCB which learns local features from convolutional activation maps, because CFCNN extracts global and multi-scale local deep features simultaneously. The performance of CFCNN is better than that of ICNN [48] by +1.9% and +2.2% improvements in rank-1 and mAP accuracy, respectively. Although ICNN also integrates global and local features, it only learns single local features, while the proposed CFCNN fuses multi-scale local features.

2) CUHK03

Table 4 shows experimental results where CFCNN obtains the highest accuracy with 64.6% and 58.4% in rank-1 and mAP, respectively. It strongly proves that jointly learning global and multi-scale local features in a unified framework is beneficial to improvement of features discriminative ability.

TABLE 3. Comparison of CFCNN with prior art on Market-1501. We categorize these methods into 3 types. The first type: global deep features; the second type: local deep features; the third type: fusion of them. We list rank-1 (%) and mAP (%) accuracy.

Methods	rank-1	mAP
V + I [47]	79.5	59.9
SVDNet [12]	82.3	62.1
PAN [18]	82.8	63.4
GatedSiamese [10]	65.9	39.6
MultiRegion [38]	66.4	41.2
PartLoss [21]	88.2	69.3
PCB [39]	92.3	77.4
SpindleNet [42]	76.9	-
MSCAN [41]	80.3	57.5
DFBP [43]	81.7	60.9
ICNN [48]	92.1	79.0
CFCNN	**94.0**	**81.2**

TABLE 4. The detailed results on CUHK03. We list rank-1 (%) and mAP (%) accuracy.

Methods	rank-1	mAP
LOMO+XQDA [6]	12.8	11.5
DeepReID [26]	19.9	-
MultiScale [49]	40.7	37.0
SVDNet+Era [50]	48.7	43.5
SI-CI [51]	52.2	-
DNS [52]	54.7	-
TriNet+Era [50]	55.5	50.7
PCB [39]	61.3	54.2
ICNN [48]	61.4	55.8
CFCNN	**64.6**	**58.4**

3) DUKEMTMC-REID

The same experiment settings are applied on DukeMTMC-reID. We list the experimental results in Table 5 where CFCNN achieves the highest values with 85.7% and 72.4% in rank-1 and mAP accuracy which demonstrate the superiority of CFCNN once again.

D. PARAMETERS ANALYSIS

In this section, four hyperparameters of the proposed CFCNN are analyzed on the Market-1501 dataset, i.e., the number of uniform horizontal stripes, the number of kernels in the convolutional layer, input pedestrian image size, and two coefficients of the CFCNN loss. It is noted that we apply the same parameters on three datasets if they are optimized.

TABLE 5. Comparison with other methods on DukeMTMC-reID. We list rank-1 (%) and mAP (%) accuracy.

Methods	rank-1	mAP
BoW+KISSME [25]	25.1	12.2
ReID+GAN [53]	67.7	47.1
OIM [54]	68.1	-
PGR [55]	72.3	52.2
AWTL [56]	79.8	63.4
DeepPerson [57]	80.9	64.8
CFCNN	**85.7**	**72.4**

TABLE 6. The influence of the number of uniform horizontal stripes for the proposed CFCNN. ✓ and ✗ indicate using and no using, respectively. Both rank-1 (%) and mAP (%) accuracy are listed.

g	n_1	n_2	dimension of D	rank-1	mAP
✓	2	6	2304	93.6	81.0
✓	3	6	2560	**94.0**	**81.2**
✓	5	6	3072	93.4	80.1
✓	6	10	4352	93.2	80.4
✓	6	15	5632	93.0	80.6
✓	6	30	9472	92.8	79.9
✓	3	✗	1024	93.3	81.0
✓	6	✗	1792	93.1	80.3
✗	3	6	2304	93.6	81.0

1) THE NUMBER OF UNIFORM HORIZONTAL STRIPES

In order to learn completed local features, we try several different combinations of n_1 and n_2, such as $n_1 = 2$ and $n_2 = 6$, $n_1 = 6$ and $n_2 = 10$ and so on. In addition, we evaluate the performance of CFCNN when only use the coarse branch and one fine branch, or just two fine branches without the coarse branch. From the 2nd row to the 7th row in Table 6, we can see that with the increase of the number of uniform horizontal stripes, the dimension of D increases higher, but the performance in rank-1 and mAP accuracy gradually declines. When using the coarse branch, $n_1 = 3$ and $n_2 = 6$, CFCNN obtains the highest accuracy in rank-1 and mAP. Furthermore, when comparing with the last three rows, i.e., fusion of global and single scale local features, and fusion of multi-scale local features without the global feature, CFCNN achieves better performance owing to jointly learning global and multi-scale local features in a unified framework.

2) THE NUMBER OF KERNELS IN THE CONVOLUTIONAL LAYER

As introduced in Section III, we apply the convolutional layer to reduce the dimension after the operation of average pooling for each branch. We vary the number of convolutional kernels from 64 to 1024. The detailed results are illustrated in Figure 6 where the proposed CFCNN achieves the best performance when the number of kernels is set to 256.

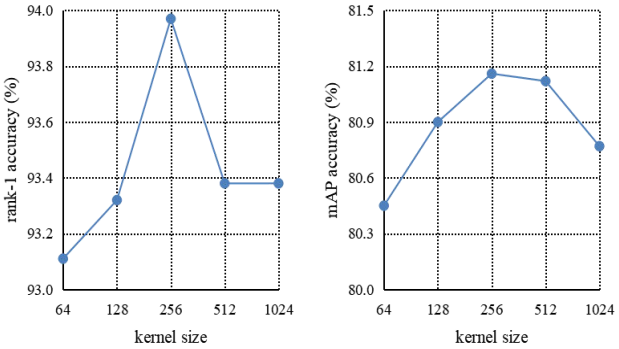

FIGURE 6. The influence of the number of convolutional kernels for the proposed CFCNN. Both rank-1 (%) and mAP (%) accuracy are listed.

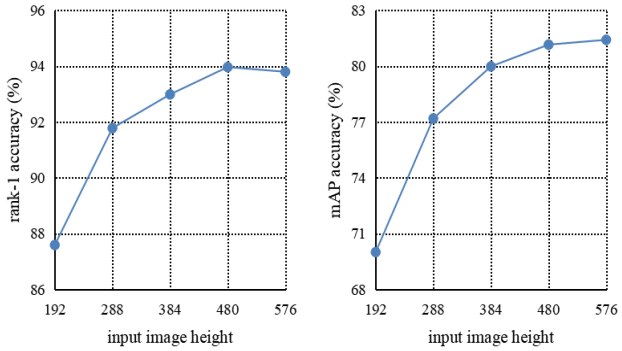

FIGURE 7. The influence of pedestrian image size for the proposed CFCNN. Both rank-1 (%) and mAP (%) accuracy are listed.

TABLE 7. The influence of λ and μ of the CFCNN loss. Only rank-1 (%) accuracy is listed.

λ \ μ	0.1	0.5	1.0	2.0	5.0
0.1	90.1	91.5	93.1	93.3	91.5
0.5	91.1	92.2	93.3	92.8	91.5
1.0	92.9	93.1	**94.0**	93.7	91.6
2.0	93.7	93.5	93.3	93.2	91.6
5.0	92.3	92.4	91.9	92.3	90.8

3) THE SIZE OF INPUT PEDESTRIAN IMAGE

We resize the pedestrian images from 192×64 to 576×192 with the aspect ratio of $3 : 1$ and report their performance in Figure 7. From Figure 7, we can see that both rank-1 and mAP accuracy show a trend of gradual increase with the pedestrian image size until reaching a stable state. Specifically, the pedestrian image with the size of 480×160 obtains almost the same performance with that of 576×192. With the consideration of memory, 480×160 image size is recommended.

4) THE COEFFICIENTS OF CFCNN LOSS

We set λ and μ with different values to control the importance of local features f_i^1 and f_j^2. The results are shown in Table 7 where CFCNN achieves the highest value with 94.0% in rank-1 accuracy when λ and μ are both set to 1.

V. CONCLUSION

In this paper, we have proposed a CNN-based model named CFCNN. The proposed CFCNN extracts global and multi-scale local features simultaneously using a multi-branch network structure. We employ the independent identity loss for each branch to enhance the discrimination of features. Finally, we obtain completed pedestrian representations via concatenating the learned global and all local features. We have proved that the proposed CFCNN has better performance than prior art on three person Re-ID datasets.

REFERENCES

[1] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1988–1995.

[2] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 3–19, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016786551200219X

[3] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. Jun. 2006, pp. 1528–1535.

[4] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2360–2367.

[5] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2011, pp. 1–11.

[6] 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.*, Jun. 2015, pp. 2197–2206.

[7] S. Tan, F. Zheng, and L. Shao, "Dense invariant feature based support vector ranking for person re-identification," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Dec. 2015, pp. 687–691.

[8] T. Matsukawa and E. Suzuki, "Person re-identification using CNN features learned from combination of attributes," in *Proc. Int. Conf. Pattern Recognit.*, Dec. 2016, pp. 2428–2433.

[9] L. Wu, C. Shen, and A. van den Hengel, "PersonNet: Person re-identification with deep convolutional neural networks," 2016, *arXiv:1601.07255*. [Online]. Available: https://arxiv.org/abs/1601.07255

[10] R. R. Varior, M. Haloi, and G. Wang, "Gated Siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 791–808.

[11] X. Zhang *et al.*, "AlignedReID: Surpassing human-level performance in person re-identification," 2017, *arXiv:1711.08184*. [Online]. Available: https://arxiv.org/abs/1711.08184

[12] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3800–3808.

[13] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3396–3405.

[14] S. Zhang, Q. Zhang, X. Wei, Y. Zhang, and Y. Xia, "Person re-identification with triplet focal loss," *IEEE Access*, vol. 6, pp. 78092–78099, 2018.

[15] Y. Liu, H. Sheng, Y. Zheng, N. Chen, W. Ke, and Z. Xiong, "GDMN: Group decision-making network for person re-identification," *IEEE Access*, vol. 6, pp. 64169–64181, 2018.

[16] S. Zhang and H. Yu, "Person re-identification by multi-camera networks for Internet of Things in smart cities," *IEEE Access*, vol. 6, pp. 76111–76117, 2018.

[17] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, Oct. 2015.

[18] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," in *Proc. IEEE Trans. Circuits Syst. Video Technol.*, to be published. doi: 10.1109/TCSVT.2018.2873599.

[19] H. Chen *et al.*, "Deep transfer learning for person re-identification," in *Proc. IEEE Int. Conf. Multimedia Big Data*, Sep. 2018, pp. 1–5.

[20] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A Siamese long short-term memory architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 135–153.

[21] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[22] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3960–3969.

[23] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," 2017, *arXiv:1701.07732*. [Online]. Available: https://arxiv.org/abs/1701.07732

[24] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 483–499.

[25] 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.*, Dec. 2015, pp. 1116–1124.

[26] 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.*, Jun. 2014, pp. 152–159.

[27] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 17–35.

[28] A. Li, Z. Wu, K. Lin, D. Chen, and G. Sun, "Self-supervised sparse coding scheme for image classification based on low rank representation," *PloS One*, vol. 13, no. 6, Jun. 2018, Art. no. e0199141.

[29] A. Li, Z. Wu, H. Lu, D. Chen, and G. Sun, "Collaborative self-regression method with nonlinear feature based on multi-task learning for image classification," *IEEE Access*, vol. 6, pp. 43513–43525, 2018.

[30] Y. Zhou, L. Liu, and L. Shao, "Vehicle re-identification by deep hidden multi-view inference," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3275–3287, Jul. 2018.

[31] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by Fisher vectors for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 413–422.

[32] Y. Hu, S. Liao, Z. Lei, D. Yi, and S. Z. Li, "Exploring structural information and fusing multiple features for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 794–799.

[33] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Comput. Vis. Image Understand.*, vol. 117, no. 2, pp. 130–144, 2013.

[34] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 536–551.

[35] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2016, pp. 1–8.

[36] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5028–5037.

[37] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1288–1296.

[38] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multi-region bilinear convolutional neural networks for person re-identification," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Aug./Sep. 2017, pp. 1–6.

[39] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 480–496.

[40] 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.*, Jun. 2016, pp. 1335–1344.

[41] 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.*, Jul. 2017, pp. 384–393.
[42] H. Zhao *et al.*, "Spindle Net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 907–915.
[43] Z. Zhang and T. Si, "Learning deep features from body and parts for person re-identification in camera networks," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, p. 52, 2018.
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available:https://arxiv.org/abs/1409.1556
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[46] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[47] 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, p. 13, Jan. 2018.
[48] Z. Zhang, T. Si, and S. Liu, "Integration convolutional neural network for person re-identification in camera networks," *IEEE Access*, vol. 6, pp. 36887–36896, 2018.
[49] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Workshops*, pp. 2590–2600, Oct. 2017.
[50] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv:1708.04896*. [Online]. Available: https://arxiv.org/abs/1708.04896
[51] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," 2017, *arXiv:1703.07220*. [Online]. Available: https://arxiv.org/abs/1703.07220
[52] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1239–1248.
[53] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3754–3762.
[54] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3415–3424.
[55] S. Liu, X. Hao, and Z. Zhang, "Pedestrian retrieval via part-based gradation regularization in sensor networks," *IEEE Access*, vol. 6, pp. 38171–38178, 2018.
[56] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6036–6046.
[57] 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