

Received August 23, 2019, accepted September 9, 2019, date of publication September 13, 2019, date of current version September 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2941002

Person Re-Identification Based on Multi-Parts of Local Feature Network

ZIMIAN WEI¹, WENJING YANG¹, WANRONG HUANG², HUADONG DAI²,
AND DONGSHENG LI¹

¹State Key Laboratory of High Performance Computing, College of Computer, National University of Defense Technology, Changsha 410073, China

²Artificial Intelligence Research Center, National Innovation Institute of Defense Technology, Beijing 100071, China

Corresponding author: Wenjing Yang (wenjing.yang@nudt.edu.cn)

This work was supported in part by the Youth Talent Lifting Project (No.17-JCJQ-QT-047), the National Natural Science Foundation of China (No. 91648204 and 61803375).

ABSTRACT Person re-identification has become increasingly popular because of its widely application in computer vision. In this paper, we propose a novel, simple and efficient person re-id network called MPLFN. The network combines two tasks: the classification task and the metric learning task. In the classification task, we uniformly partition N feature parts from an image, and compute the person classification loss in each part separately. Computing the part loss separately guides the network to focus on every body part and learn discriminative representations for each of them. And then in the metric learning task, we recalculate the distance of two images by the shortest path between two sets of feature parts. Then the distances are put into a triplet loss to perform a dynamic part alignment during the training. With the joint learning of these two tasks, the performance of the network is significantly enhanced. Compared with existing person re-id works, MPLFN achieves a better performance on three mainstream person re-identification datasets. Extensive experiments have been conducted to validate our proposed method.

INDEX TERMS Person re-identification, part loss, metric learning, alignment.

I. INTRODUCTION

Person re-identification refers to the task for retrieving specific pedestrians across different camera views in image or video sequences. In other words, it is supposed to retrieve the same person images from a gallery set, in which pedestrian images from the same camera view are excluded. There are many challenges in person re-identification, including the variability of pedestrian posture and the occlusion of camera devices with different resolutions. The prosperity of deep learning technologies (for instance, the CNN) has brought a promising solution to these challenges.

There are two widely-used CNN frameworks in person re-id: classification model [1]–[8] and metric learning model [9]–[12]. The classification model learns feature representations by minimizing the classification loss. In the test stage, given a pair of query images, the model retrieves the gallery set to compare the similarity of them. The similarity is obtained by measuring the L2 distance between learned features. The classification model can fully exploit the identity information, while the distance metrics between image pairs

are usually ignored. Differently, the metric learning model first extracts feature representations of input images and then computes the distance between them. In the training stage, the metric loss supervises the network by constraining the distance between intra-class samples and the distance between inter-class samples. The metric learning model digs similarity correlation between different images but uses weak data labels. Both classification model or metric learning model separately shows limitations in the performance of person re-id. In this paper, we propose a framework, named multi parts of local feature network (MPLFN), to combine the advantages of the classification model and the metric learning model to identify pedestrian label and learn a similarity measurement simultaneously. Specifically, the MPLFN consists of two tasks: the classification task and the metric learning task.

The target of the classification task is to learn a discriminative feature of input image by minimizing the classification loss. Considering that intuitively extract features from whole pedestrian images may ignore the discriminative local details, we conduct a uniform partition strategy and calculate the classification loss of each part separately. Note that we did not directly partition on images but equally slice the

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

globally-pooled feature representations on the height dimension. After the partition, each image can be represented by a set of feature parts. With information from other local parts filtered, each part concentrates on a saliency region. And by this means, the identity information contained in each part is fully utilized.

The target of the metric learning task is to optimize the similarity measurement of different images. We calculate the distance of different images considering the uniformly partitioned parts. In normal cases, the L2 distance of corresponding parts is calculated and added up as the distance of two samples. However, in the uniform partition cases, the corresponding parts may be not well-aligned. Therefore, we consider the distances of all possible parts by calculating a shortest path between two sets of feature parts. The distances are then put into a metric loss to supervise the training of the network.

In summary, the contributions of this paper are as follows,

- Firstly, we design a simple and efficient network named MPLFN. It's easy-to-implement and combines the classification task and metric learning task.
- Secondly, in the classification task we propose to learn local features by calculating part classification losses separately. In the metric learning task we propose to recalculate the distance of different images by considering the alignment between local parts.
- Thirdly, we implement an end-to-end person re-identification system. We conduct extensive experiments and ablation study with the widely used person re-identification datasets including Market-1501, DukeMTMC-reid and CUHK03. The results validate the effectiveness and advantages of our proposed method over existing ones.

The remainder of this paper is organized as follows: Section II presents a brief overview of recent works in person re-id. The proposed MPLFN architecture is explained in Section III. Experiment setup, results and the ablation study are presented in section IV. And the conclusions are given in Section V.

II. RELATED WORK

In recent years, significant improvement has been achieved in person re-id with the assistance of deep learning. In general, person re-id methods are usually based on two types of models: the classification model and the metric learning model.

The classification model [1]–[8] mainly aims at learning a discriminative feature representation of input image. Reference [1] proposes a novel dropout strategy to learn features from multiple datasets jointly by classification loss. Reference [2] trains the network jointly with identification and center losses [13]. References [3]–[8] learn part-based local feature in classification model. References [3], [4] apply pose estimation methods to crop the human region precisely. References [5], [6] adopt human semantic parsing method to help filter background clutter and occlusion. Different local

features are fused and put into a classification loss to enhance the discriminative capacity. However, the transfer deviations between different datasets impeded the performance of these methods. References [7], [8] abandon human part annotations and apply part losses to guide the network learn local features automatically. However, they did not cover the similarity correlation between different samples. In our proposed method, we apply multi part loss as in [7], [8] yet joint learning with a metric learning task.

In terms of metric learning model [9]–[12], [14], it first extracts feature representations of input images, then computes the distance of feature representations as similarities. Afterwards the distances are put into a metric learning loss. While training, metric learning loss supervises the network to pull intra-class samples closer and push inter-class samples further. The input of metric learning model is usually in pairs (intra-class pairs and inter-class pairs) for Siamese network or in triplet set (anchor, positive and negative) for triplet network. References [9], [10] propose a Siamese network with contrastive loss that constrain the inter-class distance is larger than intra-class distance by a margin. Reference [11] proposes a novel triplet network with multi-channel feature fusion. The features are then put into a triplet loss to pull the distance of positive samples closer and at the same time push the distance of negative samples further. Reference [12] proposes an advanced batch hard sampling method in triplet loss to facilitate effective learning. In our proposed method, in metric learning task we apply the batch hard triplet loss in [12], the modification is that we recalculated the distance considering the local alignment of uniformly partitioned strips. Reference [14] proposes a matrix metric learning method. The feature description is transformed from characteristic vector to discrepancy matrix. However, the performance is vulnerable with the reference sets. The hand-craft objective function in [14] may fails to learn a high-dimensional information from the input data. In our proposed method, we modified the pre-trained resnet101 model to extract discriminative deep features and then put the divided part features into metric learning task and classification task. It's an end-to-end framework and the deep features are more robust and discriminative compared with hand-crafted features.

There are some works that integrate classification and metric learning models [15]–[19]. Specifically, [15] combines classification loss, triplet loss and center loss. A set of training tricks are applied in [15] to improve the performance. For example, learning rate warming up, random erasing, label smoothing and so on. Generally, our method is different with [15] as follows: our method learns from local feature while [15] learns from global feature. Our classification loss is summed up by part losses. In our metric learning loss, the distances of positive pairs and negative pairs are recalculated by considering the consistency between two sets of local parts. In [15], the features are extracted from the whole image, which may ignore the discriminative local details. Moreover, our method reports experiments on mainstream person re-id dataset CUHK03 [20], which is not included in [15].

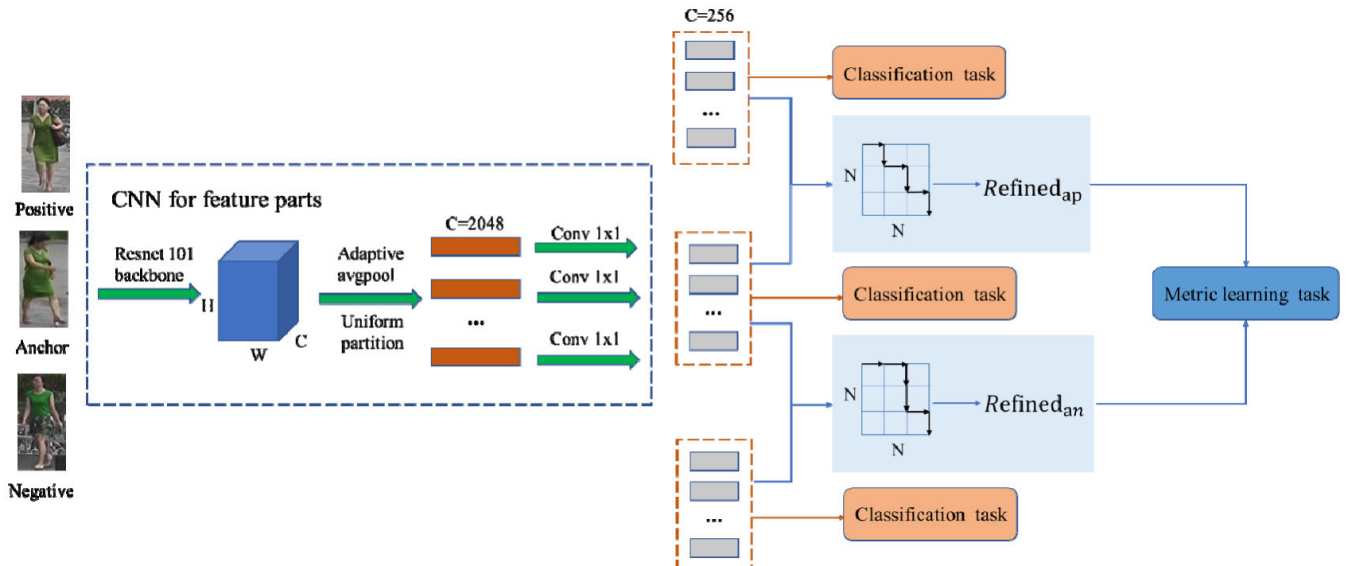


FIGURE 1. The framework of multi-parts of local feature network. Anchor is an image that randomly sampled in the training set. Positive is the same identity with anchor, while negative is not. The module “CNN for feature parts” are detailed in Table 1. The module aims to extract a discriminative feature tensor and then uniformly partition the feature tensor for the classification task and the metric learning task. In “CNN for feature parts”, the input image goes through resnet101 backbone network to form a 3D feature tensor ($C \times H \times W$). The adaptive avgpool is conducted to obtain designated spatial size ($H \times W$) of the feature tensor. We uniformly partition the feature tensor on the height dimension (H). The conv 1x1 reduce the feature channel (C) from 2048 to 256. One input image can be represented by a set of feature parts. Then the model combine with two tasks: the classification task and the metric learning task, which are detailed in Fig. 2. The Refined_{ap} and Refined_{an} are recalculated distances of the positive pairs and negative pairs. They are based on the shortest path in the $N \times N$ feature distance matrix. N is the number of feature parts of each image. The aim of local alignment method is to recalculate the distance between positive pairs and negative pairs, the distances are then put into a triplet loss for a metric learning task.

References [16]–[19] target at unsupervised person re-id, which is a new type of re-id model. Reference [16] is the first one using unsupervised method for person re-id. Reference [17] is a classification model, but it leverages the metric information during its training. Its improved version [18] combines the classification loss and an exclusive loss (which is a metric learning loss in a classification way). Reference [19] introduces clustering on the exclusive loss. Unsupervised person re-id methods have a promising prospect but the performance show limitations in the current stage. In this paper, we focus on supervised person re-id and regard the unsupervised person re-id as a future work.

Reference [21] attempts to promote the re-id effectiveness by an additional re-ranking step. It optimizes the ranking list by accumulating historical queries. Reference [21] targets at the practical person re-id process but did not develop core re-identification model technology. In our proposed method, we propose a simple and efficient person re-id model. It's easy to implement and the extensive experiments have proved the effectiveness and advantages over existing methods.

III. PROPOSED METHOD

In this section, we present the proposed MPLFN in details. First of all, we describe the architecture of MPLFN. Then, we introduce the losses of classification task and metric learning task.

A. ARCHITECTURE

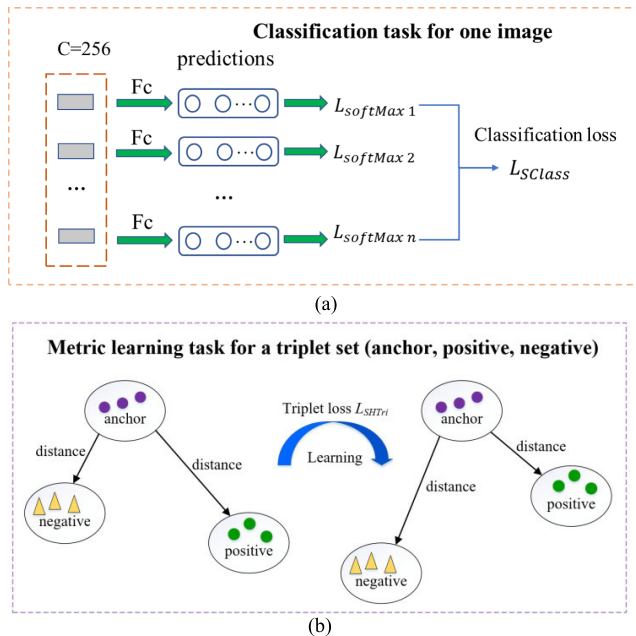
The architecture of the proposed method is shown in Fig. 1 and Fig. 2. We introduce the details of “CNN for

feature parts” in Table 1. Considering both computation cost and performance, we employ the pre-trained ResNet-101 [22] model with several modifications as part of our backbone network. The modifications we do with ResNet-101 are as follows: First of all, we enrich the feature granularity by setting all stride of down-sampling and convolution layer in conv5_x block to 1. Then the last two layers (average pool layer and fully connected layer) are removed to obtain the feature representations from the last convolutional layer. The image sample comes down the backbone layers will be transformed to a 3D feature tensor with the shape of $C \times H \times W$, C is the feature channel ($C = 2048$) and $H \times W$ is the spatial size. Afterwards, the adaptive average pool is operated on the feature tensor to obtain designated spatial size. Then we conduct a uniform partition along the tensor height. The feature tensor is divided into several horizontal parts. A conv 1d block is designed to reduce the feature channel of divided feature parts from 2048-dim to 256-dim. Finally we get a set of feature parts. Note that the number of feature parts is not fixed. In this paper, we set the number as 3. Different number of parts are compared in section IV-D. We choose from the best experiment performance.

We combine classification task and metric learning task on feature parts. For the classification task, each feature part of one image is separately put into a classifier. The classifier contains a fully connection layer and a SoftMax layer. Then the classification loss of each part is calculated and summed up. For the metric learning task, we apply a local

TABLE 1. Details of “CNN for feature parts” module in Fig. 1. Images that go through the “CNN for feature parts” are represented by a set of feature parts.

Name	Kernel size	stride	padding	Output feature size(C×H×W)
Image input				3×384×192
Conv1	7×7, 64	2	(3,3)	64×192×96
Maxpool	3×3	2	(1,1)	64×96×48
Conv2_x	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \times 3$	$\begin{bmatrix} (0,0) \\ (1,1) \\ (0,0) \end{bmatrix} \times 3$	256×96×48
Conv3_x	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \times 3$	$\begin{bmatrix} (0,0) \\ (1,1) \\ (0,0) \end{bmatrix} \times 4$	512×48×24
Conv4_x	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \times 22$	$\begin{bmatrix} (0,0) \\ (1,1) \\ (0,0) \end{bmatrix} \times 23$	1024×24×12
Conv5_x	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \times 3$	$\begin{bmatrix} (0,0) \\ (1,1) \\ (0,0) \end{bmatrix} \times 3$	2048×24×12
Adaptive avgpool2d(3,1)				2048×3×1
uniform partition				2048×1
				2048×1
				2048×1
Conv 1d	[1,256]	1	-	256×1 256×1 256×1

**FIGURE 2.** The classification task and the metric learning task in MPLFN. (a) describe the classification task for one image. Each feature part of one image are put into a classifier separately. The classifier consists of a fully connection layer and a SoftMax layer. Then the classification loss of each part are calculated and summed up to form the classification loss of MPLFN. (b) describe the metric learning task for a triplet set including anchor, positive, and negative images. The distances of positive pairs and negative pairs are feed into a metric loss to supervise the training of the network.

alignment method to obtain the distance between two sets of feature parts. Then the distances of a triplet samples (distance between positive pairs and distance between negative pairs)

are feed to a triplet loss to learn the similarity of different images. The classification loss and the metric learning loss are combined to supervise the training of the network. In test stage, the feature parts are concatenated to compute the rank results

B. LOSS FUNCTION

We integrate classification loss with metric learning loss while training to maximize the discrimination ability of the proposed method. The loss of the proposed method can be formulated as:

$$L = L_{SCl} + L_{SHTri} \quad (1)$$

where L_{SCl} is the classification loss, L_{SHTri} is the metric learning loss.

1) CLASSIFICATION LOSS

The target of classification loss is to supervise the network to learn identity-specific features. After an image is represented by a set of feature parts, we consider the identification task as a multi-classification problem.

For i -th feature parts f_i , the classification loss is formulated as:

$$L_{f_i} = - \sum_{i=1}^M \log \frac{e^{W^T f_i + b}}{\sum_{i=1}^C e^{W^T f_i + b}} \quad (2)$$

where W, b denote the parameter of the fully connected layer, C denotes the number of classes in the training set M is the size of mini-batch in training process.

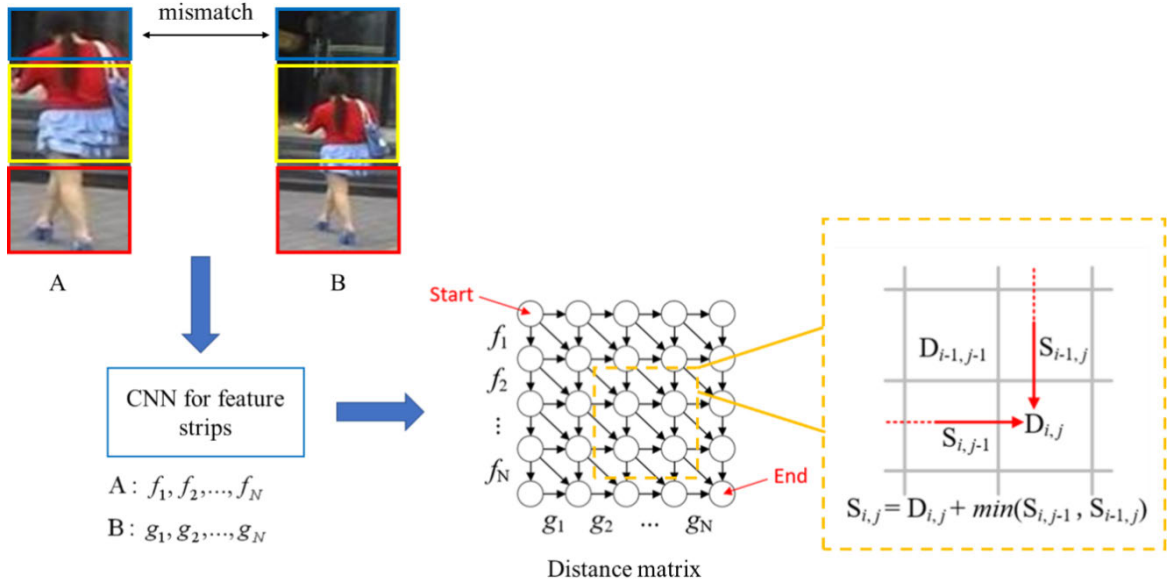


FIGURE 3. Image A and B are with the same identity. In “CNN for feature parts”, the feature tensors of A and B are uniformly partitioned on the height dimension. The mismatch on the horizontal direction of image A and B may impede the precisely distance measurement. In metric learning task, we apply a dynamic local alignment method [15] to refine the distance of two images. The shortest path from the start to the end in the feature distance matrix is returned as the refined distance between image A and B.

The classification loss of the proposed method is the summation of part classification losses:

$$L_{\text{SClass}} = \sum_{i=1}^N L_{f_i} \quad (3)$$

where N is the number of parts. In this paper, we set $N = 3$. The discussion of N is in section IV-D.

2) METRIC LEARNING LOSS

The metric learning loss of the proposed method consists of two components. Firstly, we refine the similarity distance by applying an dynamic local alignment method [23]. Afterwards, the refined distances of positive pairs and negative pairs are feed into a triplet loss to learn the similarity of different images. Details are as follows:

a: REFINED DISTANCE

As shown in Fig. 3, image A and B are with the same identity. In “CNN for feature parts” module, the feature tensor of them are uniformly partitioned along the height dimension. However, the corresponding feature parts of A and B may be mismatched, which will lead to a wrong measurement of similarity distance. In metric learning task, we refine the distance of different images as follows:

Given two images, they are represented by two sets of feature parts. $X = f_1, f_2, \dots, f_N$, $Y = g_1, g_2, \dots, g_N$. N is the number of parts in each sets.

Firstly, we build a distance matrix $D_{N \times N}$ of feature parts. $D_{i,j}$ is the L_2 distance between f_i and g_j

Secondly, elements in $D_{N \times N}$ is normalized to avoid over-fitting. “-1” and “+1” is used to normalize the element to $[0,1)$

$$D_{i,j} = \frac{\exp(D_{i,j}) - 1}{\exp(D_{i,j}) + 1}, \quad i, j \in 1, 2, \dots, n \quad (4)$$

Thirdly, in distance matrix, we use the shortest path from start to end ($S_{N,N}$) as the similarity distance of two images A and B. The shortest path $S_{i,j}$ is calculated as follows:

$$S_{i,j} = \begin{cases} D_{i,j} & i = 1, \quad j = 1 \\ S_{i-1,j} + D_{i,j} & i \neq 1, \quad j = 1 \\ S_{i,j-1} + D_{i,j} & i = 1, \quad j \neq 1 \\ \min(S_{i-1,j}, S_{i,j-1}) + D_{i,j} & i \neq 1, \quad j \neq 1 \end{cases} \quad (5)$$

$S_{N,N}$ is returned as the refined distance of two images. Details are shown in Alg.1.

b: Function formulation

We apply the effective batch-hard sampling method in [12] to mine candidate triplet sets. P classes are randomly sampled in batch and K images are randomly sampled in each class. We set $P = 16$ and $K = 4$ in our proposed method. For a randomly sampled anchor, the positive sample with the largest distance and the negative sample with the smallest distance are selected to construct a candidate triplet.

The candidate triplet samples are stored in M . The metric loss of the proposed method can be formulated as follows:

$$L_{\text{SHTri}} = \sum_{(a,p,n) \in M} [\alpha + \text{refined}_{ap} - \text{refined}_{an}]_+ \quad (6)$$

Algorithm 1 Algorithm of Local Alignment

Input: Two sets of feature parts $(f_i)_{i=1:N}$, $(g_j)_{j=1:N}$
Output: Refined distance $S_{N,N}$ after local alignment

```

1  for  $i = 1 : N$  do
2    for  $j = 1 : N$  do
3       $D_{i,j} \leftarrow [\exp(\|f_i - g_j\|_2) - 1] / [\exp(\|f_i - g_j\|_2) + 1]$ 
4    end for
5  end for
6  for  $i = 1 : N$  do
7    for  $j = 1 : N$  do
8      if  $i == 1$  and  $j == 1$  then
9         $S_{i,j} \leftarrow D_{i,j}$ 
10     else if  $i == 1$  and  $j > 1$  then
11        $S_{i,j} \leftarrow D_{i,j} + S_{i,j-1}$ 
12     else if  $j == 1$  and  $i > 1$  then
13        $S_{i,j} \leftarrow D_{i,j} + S_{i-1,j}$ 
14     else
15        $S_{i,j} \leftarrow \min(S_{i-1,j}, S_{i,j-1}) + D_{i,j}$ 
16     end if
17   end for
18 end for
19 Return  $S_{N,N}$ ;

```

In (6), α is a margin parameter set as 0.3. M is constructed by batch-hard sampling method in [12]. (a, p, n) present for anchor, positive, and negative samples. $refined_{ap}$ is the refined distance of anchor and positive samples. $refined_{an}$ is the refined distance of anchor and negative samples. By minimizing the metric learning loss, the intra-class samples are getting closer and the inter-class samples are pushed away. The discriminative capacity of the network is enhanced.

IV. EXPERIMENTS

In this section, we show the experiment details of the proposed MPLFN.

A. EXPERIMENT SETUP

Each image is resized to 384×192 for obtaining detailed information from person images. Normalization, shuffle, and random horizontal flip are used as data augmentation. The backbone model is initialized by weights of ResNet-101 pre-trained on ImageNet [24]. The batch size is set to 32. While in the training phase, we use SGD optimizer with momentum 0.9. The weight decay is $5e-4$. We set the initial learning rate as 0.1. Specially, learning rate in backbone network is reduced by 0.1. At the same time, we decay learning rate every 20 epoch, attenuation rate is 0.1. The total training process is lasted for 100 epochs. Our model is implemented on PyTorch platform. With NVIDIA GeForce GTX1070 GPU, it takes about 11 hours for training on Market-1501. Experiments on different datasets follow the same settings above.

TABLE 2. Training/testing protocols for three datasets. In A/B, A means identity, B means images.

datasets	train	query	gallery
Market-1501	751/12936	750/3368	751/15913
DukeMTMC-reID	702/16522	702/2228	1110/17661
CUHK03	767/7365	700/1400	700/5332

TABLE 3. Comparison with other methods on Market-1501 dataset.

Methods	R-1	mAP
SVDNet ^[28]	82.3	62.1
PAN ^[29]	82.8	63.3
FMN ^[30]	86.0	67.1
DPFL ^[31]	88.9	73.1
GLAD ^[32]	89.9	73.9
MLFN ^[33]	90.0	74.3
HA-CNN ^[34]	91.2	75.7
Bilinear ^[35]	91.7	79.6
Mancs ^[36]	93.1	82.3
DeepCRF ^[37]	93.5	81.6
PCB+RPP ^[8]	93.8	81.6
MPLFN (Ours)	94.2	82.4

B. DATASETS AND PROTOCOLS

Our proposed method is evaluated on three widely used person re-identification datasets: Market-1501 [25], CUHK-03 [20] and DukeMTMC-reID [26]. The Market-1501 dataset contains 32,668 annotated bounding boxes with 1501 identities. Each identity contains about 17 images. The images of each identity are captured by up to six cameras. The DukeMTMC-reID dataset contains 36,411 images of 1812 identities. Images are captured by 8 high-resolution cameras. The hand-crafted person attributes are provided. The CUHK03 dataset contains 14,097 images with 1467 identities. The images of each identity is captured by two cameras. Both bounding boxes detected from deformable part models (DPMs) and manual labeling are provided for annotations.

We follow the official training and evaluation protocols on Market-1501 and DukeMTMC-reID. As for CUHK03, we adopt the new training/testing protocol proposed in [27] since the original protocol is time consuming. Details of training/testing protocols are shown in Table 2. The cumulative matching characteristics (CMC) at rank-1 and mean average precision (mAP) are used for evaluating the performance of the proposed method. Note that all the experiments we conducted are in single-query mode.

C. RESULTS AND ANALYSIS

We conduct comparative experiments with current state-of-art methods to show the superiority of our proposed method. The results in detail are as follows. ‘-’ means that experimental data not given in the original paper.

The comparison results of the proposed MPLFN method and state-of-art methods on Market-1501 dataset are reported in Table 3. In single-query mode, [8] have reported best published performance with mAP = 81.6%/ Rank-1=93.8%.



FIGURE 4. Top 10 ranking list by MPLFN for some query images on Market-1501 datasets. The first column are query images. The following 10 columns are ranked results retrieved by MPLFN. The red border indicates a wrong match.

TABLE 4. Comparison with other methods on DukeMTMC-reID dataset.

Methods	R-1	mAP
GAN ^[38]	67.7	47.1
PAN ^[29]	71.6	51.5
FMN ^[30]	74.5	56.9
SVDNet ^[28]	76.7	56.8
HA-CNN ^[34]	80.5	63.8
Bilinear ^[35]	82.1	64.2
Mancs ^[36]	84.9	71.8
DeepCRF ^[37]	84.9	69.5
PCB+RPP ^[8]	83.3	69.2
MPLFN (Ours)	85.7	72.6

However, Our MPLFN method achieves $mAP = 82.4\%$ and Rank-1 = 94.2% without re-ranking [27], pulling the baseline of person re-id on Market-1501 to a new level.

Fig. 4 shows top-10 ranking results by MPLFN for some given query images on Market-1501 datasets. For each query image in the first column, the proposed MPLFN method retrieves top-10 ranking results in the gallery set. Images with red borders indicates a wrong shot. Images without red borders means the retrieved images are of the same identity with the query image. We can observe from the visualization

TABLE 5. Comparison with other methods on CUHK03 dataset. We apply new training/testing protocol proposed in [20] since the original version is time consuming.

Settings	767/700 split			
	Labeled		Detected	
	R-1	mAP	R-1	mAP
SVDNet ^[28]	40.9	37.8	41.5	37.3
FMN ^[30]	40.7	38.0	42.6	39.2
DPFL ^[31]	43.0	40.5	40.7	37.0
Pose-Transfer ^[39]	45.1	42.0	41.6	38.7
HA-CNN ^[34]	44.4	41.0	41.7	38.6
MLFN ^[33]	54.7	49.2	52.8	47.8
PCB+RPP ^[8]	-	-	63.7	57.5
MPLFN (Ours)	67.9	64.8	68.4	65.1

that the proposed MPLFN method shows great robustness and high accuracy in person re-identification.

In Table 4, we compare our proposed MPLFN method with state-of-art methods on DukeMTMC-reID dataset. Our method achieves $mAP = 72.6\%$ and Rank-1 = 85.7%, which has proved the excellent performance of MPLFN.

The CUHK03 dataset is divided into machine labeled (detected) and manually labeled (labeled). Table 5 lists the comparison results using two different annotations. Note that

TABLE 6. Ablation study of MPLFN on three datasets. Resnet101 backbone represents MPLFN method with a single classification loss. Resnet101+MS represents MPLFN method with multi-part classification loss. Resnet101+MS+SD represents MPLFN with both multi-part classification loss and local aligned metric learning loss.

Model	Market-1501		DukeMTMC-reID		CUHK03 (detected)	
	R-1	mAP	R-1	mAP	R-1	mAP
Resnet101 backbone	90.6	78.4	82.6	67.3	59.6	57.4
Resnet101+MS	93.4	81.3	84.9	70.6	65.6	63.1
Resnet101+MS+SD(MPLFN)	94.2	82.4	85.7	72.6	68.4	65.1

we follow new 767/700 split proposed in [20] since the original version is time consuming. Our MPLFN method achieves mAP = 64.8% and Rank-1 = 67.9% for labeled annotation while mAP = 65.1% and Rank-1 = 68.4% for detected annotation, which surpasses state-of-art methods by a large margin.

D. COMPONENT ANALYSIS

In this section, we separately analysis important components in MPLFN. Moreover, we design several alternative models to compare with MPLFN.

1) IMPACT OF MULTI-PART CLASSIFICATION LOSS

We compare the experimental results of resnet101 backbone with single loss, and MPLFN with classification loss to analysis the impact of multi-part classification losses. There are no feature parts and metric learning loss in resnet101 backbone, so we just use the multi-part classification losses in MPLFN. We use Resnet101 backbone to represent MPLFN with single classification loss. Note that Resnet101 backbone is different with resnet101 in the stride of Conv5_x block. We use Resnet101+MS to represent MPLFN with multi-part classification losses. We can observe significant performance improvement on Table 6. The results proved that the multi-part losses can improve the performance of single loss.

2) IMPACT OF METRIC LEARNING TASK

We conduct comparative experiments to prove the effect of metric learning task on MPLFN. As detailed in section III-B, the local alignment algorithm refined the distance of different images, then they are put into a metric learning loss to supervise the training of the network. We use Resnet101+MS to represent MPLFN method just with classification loss. We use Resnet101+MS+SD to represent MPLFN with both classification loss and local aligned metric learning loss. As we can see from Table 6 and Fig. 5, the resnet101+MS+SD have further improved the performance of resnet101+MS.

3) IMPACT OF FEATURE PARTS' NUMBER

We vary the feature parts' number N in MPLFN to see the impact of the feature parts' number. Results are shown in Fig. 5. When $N = 1$, the MPLFN learns with a single classification loss. The results show that as N increases, the performance improved at first until $N = 3$. When N increases to more than 3, as N get larger, the performance

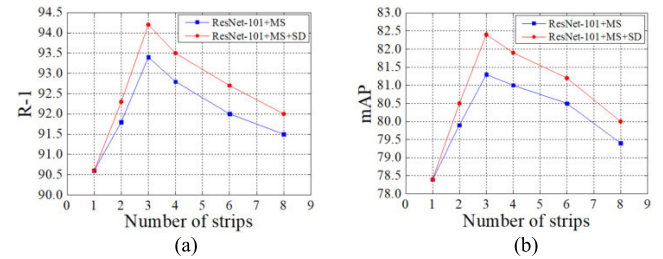


FIGURE 5. Impact of part number on MPLFN. (a) Results of rank-1 accuracy on market-1501 dataset. (b) Results of mAP on market-1501 dataset.

TABLE 7. Comparison with alternative structures on Market-1501.

Model	Market-1501	
	R-1	mAP
Feature-2048	93.2	81.0
Shared parameter	93.0	80.4
Mid-fused feature	91.5	78.6
MPLFN	94.2	82.4

begin to drop. We attribute the reason to the too large granularity of part features, which may lead to overlap in local features. So in this paper, we choose $N = 3$.

4) ALTERNATIVE STRUCTURE

We design several alternative structures to compare with MPLFN to comprehensively analysis the effectiveness of MPLFN.

Feature-2048: In MPLFN, the feature parts are reduced to 256-dim by a conv1x1 block before put into a classification task. The Feature 2048 is the alternative model without the dimension reduction.

Shared parameter: In MPLFN, each feature part do a classification task separately. In **Shared parameter** model identity classifiers shared parameters with each other.

Mid-fused feature: In MPLFN, the feature is extracted from the last convolution layer. Inspired by [40], **Mid-fused Feature** model conduct a feature fusion before uniform partition. The output mid-feature from last three convolution layers are integrated with the last-layer feature. The rest of structure stay the same with MPLFN.

We infer from the experimental results in Table 7 that the performance of **Feature-2048** is slightly slower than MPLFN, with R-1/mAP= -1.0%/-1.4%. The shared

parameter between identity classifiers in **Shared parameter** brings performance decay by $R-1/mAP = -1.2\%/-2.4\%$. It indicates that shared parameters is not favorable for exploiting the identity information located in each local parts. **Mid-fused feature** integrates features extracted from the last 3 convolutional layers introduced a dropped performance of $R-1/mAP = -2.7\%/-3.8\%$. It shows that the last convolutional layer contains more high-level semantic features and is more suitable for uniform partition.

5) MODEL COMPLEXITY

We compare the model complexity of MPLFN and baseline to discuss the effectiveness of our proposed method. As mentioned in section IV-A, MPLFN takes about 11 hours to train on Market-1501 with a NVIDIA GeForce GTX1070 GPU. The ResNet101 baseline is MPLFN with a single classification loss. Under the same settings, the time cost of ResNet101 baseline is about 9 hours. Since the time cost can be influenced by different training tricks, the model complexity is usually represented by the size of model parameters. The model size of ResNet101 baseline is about 43M. The model size of MPLFN is about 46M. We can infer from Fig. 5 that when compared with ResNet101 baseline ($N=1$), our method achieves the performance of $+3.6\%$ in Rank-1 and $+4.0\%$ in mAP with slightly larger model complexity. There are also person re-id works that concentrate on designing lightweight networks, for example [41]. Compared with these methods, the model complexity has become a drawback of our proposed method. In the future work, we will focus on decreasing the model complexity to improve our proposed method.

V. CONCLUSION

In this paper, we propose a novel person re-id method based on multi-parts of local feature network. The proposed method combine classification task and metric learning task to learn the discriminative features and an effective similarity measurement simultaneously. In classification task we conduct a uniform partition strategy and calculate part classification losses separately to fully exploit the identity information contained in each part. In metric learning task, we recalculate the distance of image pairs by considering the alignment between the two sets of parts. The distances are then put into a triplet loss. Both classification loss and triplet loss are applied to supervise the training of the network. Extensive experiments on Market-1501, DukeMTMC-reid and CUHK03 have proved that our proposed method have pushed the performance of person re-id to a new level. Future works will focus on decreasing the model complexity and applying our method to the unsupervised person re-id field.

REFERENCES

- [1] 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)*, Las Vegas, NV, USA, Jun. 2016, pp. 1249–1258.
- [2] H. Jin, X. Wang, S. Liao, and S. Z. Li, "Deep person re-identification with improved embedding and efficient training," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Denver, CO, USA, Oct. 2017, pp. 261–267.
- [3] 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. (ICCV)*, Venice, Italy, Oct. 2017, pp. 3980–3989.
- [4] 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)*, Honolulu, HI, USA, Jul. 2017, pp. 907–915.
- [5] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1062–1071.
- [6] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang, "Eliminating background-bias for robust person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5794–5803.
- [7] 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.
- [8] 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. (ECCV)*, 2018, pp. 480–496.
- [9] R. R. Viorio, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2016, pp. 791–808.
- [10] R. R. Viorio, 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.*, 2016, pp. 135–153.
- [11] 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)*, Las Vegas, NV, USA, Jun. 2016, pp. 1335–1344.
- [12] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," Mar. 2017, *arXiv:1703.07737*. [Online]. Available: <https://arxiv.org/abs/1703.07737>
- [13] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.
- [14] 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. doi: [10.1109/TCYB.2017.2755044](https://doi.org/10.1109/TCYB.2017.2755044).
- [15] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019.
- [16] 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, p. 83, Oct. 2018.
- [17] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by step-wise learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5177–5186.
- [18] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2872–2881, Jun. 2019.
- [19] 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.*, vol. 2, 2019, pp. 1–8.
- [20] 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.*, Columbus, OH, USA, Jun. 2014, pp. 152–159.
- [21] Z. Wang, J. Jiang, Y. Yu, and S. Satoh, "Incremental re-identification by cross-direction and cross-ranking adaption," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2376–2386, Sep. 2019.
- [22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
- [23] 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," Nov. 2017, *arXiv:1711.08184*. [Online]. Available: <https://arxiv.org/abs/1711.08184>
- [24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.

- [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. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1116–1124.
- [26] 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.* Cham, Switzerland: Springer, 2016, pp. 17–35.
- [27] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3652–3661.
- [28] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 3820–3828.
- [29] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
- [30] G. Ding, S. Khan, Z. Tang, and F. Porikli, "Let features decide for themselves: Feature mask network for person re-identification," Nov. 2017, *arXiv:1711.07155*. [Online]. Available: <https://arxiv.org/abs/1711.07155>
- [31] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Venice, Italy, Oct. 2017, pp. 2590–2600.
- [32] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-local alignment descriptor for scalable person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 986–999, Apr. 2019.
- [33] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2109–2118.
- [34] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2285–2294.
- [35] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 402–419.
- [36] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Manacs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 365–381.
- [37] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep CRF for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8649–8658.
- [38] 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. (ICCV)*, Venice, Italy, Oct. 2017, pp. 3774–3782.
- [39] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4099–4108.
- [40] Q. Yu, X. Chang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching," Nov. 2017, *arXiv:1711.08106*. [Online]. Available: <https://arxiv.org/abs/1711.08106>
- [41] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," May 2019, *arXiv:1905.00953*. [Online]. Available: <https://arxiv.org/abs/1905.00953>



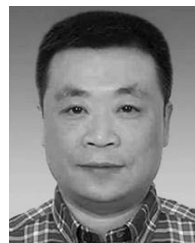
ZIMIAN WEI received the M.S. degree in software engineering from the National University of Defense Technology, in 2018, where she is currently pursuing the Ph.D. degree. Her main research interests include computer vision, machine learning, and reinforcement learning.



WENJING YANG received the Ph.D. degree in multi-scale modeling from Manchester University. She is currently an Associate Research Fellow with the State Key Laboratory of High Performance Computing, National University of Defense Technology. Her research interests include machine learning, robotics software, and high-performance computing.



WANRONG HUANG received the Ph.D. degree in computer science from the National University of Defense Technology (NUDT), China, in 2018. He is currently an Assistant Research Fellow with the National Innovation Institute of Defense Technology, Beijing, China. His research interests include the distributed control of multi-robot systems, computer architecture, and artificial intelligence.



HUADONG DAI received the Ph.D. degree from the National University of Defense Technology, China, in 2002. He is currently a Professor with the National Innovation Institute of Defense Technology, Beijing, China. His main research interests include operating systems, computer architecture, and artificial intelligence.



DONGSHENG LI received the B.Sc. and Ph.D. degrees (Hons.) in computer science from the College of Computer Science, National University of Defense Technology, Changsha, China, in 1999 and 2005, respectively. He is currently a Full Professor with the National Laboratory for Parallel and Distributed Processing, National University of Defense Technology. His research interests include parallel and distributed computing, cloud computing, and large-scale data management. He was awarded the prize of National Excellent Doctoral Dissertation of PR China by the Ministry of Education of China, in 2008.

...