IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 12, 2020, accepted December 21, 2020, date of publication December 30, 2020,
date of current version January 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3048119

Multi-Stream Refining Network for Person Re-Identification

XU WANG[1,2], YAN HUANG[1,2], QICONG WANG[1,2], YAN CHEN[3], AND YEHU SHEN[4]

[1]Shenzhen Research Institute, Xiamen University, Shenzhen 518000, China
[2]Department of Computer Science, Xiamen University, Xiamen 361005, China
[3]College of Business and Management, Xiamen Huaxia University, Xiamen 361021, China
[4]School of Mechanical Engineering, Suzhou University of Science and Technology, Suzhou 215009, China

Corresponding author: Qicong Wang (qcwang@xmu.edu.cn)

This work was supported in part by the Shenzhen Science and Technology Projects under Grant JCYJ20180306173210774 and Grant JCYJ20200109143035495, and in part by NSFC under Grant 51975394.

ABSTRACT Viewpoint change, pose variation and background clutter have adverse impacts on similarity evaluation for person re-identification. Because of its distinction and reliability, person saliency has been applied to model person appearance characteristics. However, such valuable information is not fully exploited to compute similarities of person images with existing deep methods. To this end, we present a novel multi-stream refining based deep multi-task learning scheme that aggregates multi-stage salient embedding features in the network to boost the retrieval performance. Specifically, the backbone network is divided into four stages and a channel significance self-learning sub-module is introduced to strengthen the importance of saliency channels adaptively. Meanwhile, an enhancement sub-module is employed to extract the common information and different information from the channels. Finally, a multi-stream multi-task learning framework combining four-stage branches is adopted to learn discriminative features. Compared with the state-of-the-art approaches, our model achieves competitive performance on three publicly available datasets, i.e., Market-1501, MSMT17, and CUHK03. The experimental results demonstrate the superiority of our method, which achieves 95.67%/88.51%, 87.53%/65.54%, and 89.32%/78.99% on Rank-1/mAP, respectively.

INDEX TERMS Salient channels, refining module, multi-stream, multi-task.

I. INTRODUCTION

Person re-identification (re-ID), receiving more and more attention, is one of the most active topics in computer vision [1]. It can be widely applied in intelligent surveillance, video analysis, and other fields. Generally speaking, the re-ID model is to use machine learning techniques to identify the same person in non-overlapping camera views. However, present approaches [1]–[7] mainly suffer from the following limitations.

1) Low resolution. Because a large number of people's images are taken by cameras, most of them have low resolution. Moreover, the distance between the camera and the object is generally quite far, so the resolution of the person in the image is relatively low.

2) Variations of viewpoints and poses. Generally speaking, the camera captures the subjects randomly from different viewpoints, so people in the images often present a variety of poses and viewpoints.

3) Illumination changes. Different camera positions and acquisition times may cause large variants in brightness, which may change the visual appearance of the people and affect the performance of the model.

4) Background and occlusion. The re-ID model is usually used to identify images taken from cameras with non-overlapping views. Complex background and occlusions will lead to a lot of noise in the extracted features, which will seriously affect the accuracy of re-ID.

At present, person re-ID models based on the deep networks are one of the research hotspots [6]. In deep feature space, the channels in the feature map are closely related to the rich information in the image. Some channels can represent significant information about the image, which are called salient channels, which is shown in Figure 1. For example, in Fig. 1 (a), a person's red shirt is a significant area that can be used to distinguish identities, so its corresponding

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

channels are salient. This is on channel 8. It can be seen that the channel contains only significant information corresponding to the red shirt area. Figure 1 (b) is the visualization of channel 36, which can roughly represent the important information of the whole human body area. Therefore, increasing the weights (significances) of these representative channels can help to reduce the influence of redundant background and partial occlusion. In Fig. 1 (c), the sixth channel is visualized, and it represents the salient texture information, which is superior to the color information in the case of strong illumination change. For the channels shown in Fig. 1 (d) and (e), they mainly correspond to some background regions. If they have low responses, the re-ID model can reduce the interference of these channels. It can be seen from Figure 1 that some salient channels are discriminative. If their weights can be continuously strengthened during network training, it will be very helpful for the re-ID model to deal with some constraints effectively, such as background clutter and viewpoint changes.

Salient channels also reflect view of invariance. As shown in Fig. 2, P1, P2, P3, and P4 are four pairs of persons with different identities. (a) and (b) represent two image sets of the same persons taken from different camera viewpoints. It can be intuitively observed that the salient area of P1 is in the upper body, in P2 it is on the skirt, in P3 it is the whole body, and in P4 it is a red package. The salient information is nearly the same under different viewpoints. These features have an important contribution to the recognition of the different identity of the person. From Figure 2, we find that the importance of the information contained in the salient channels are unequal. Compared with the channel information of other people, the area corresponding to the red packet in the salient channel of person P4 is more important and distinctive. Besides, it is the common characteristic for the person as P4.

However, most network structures [8]–[14] treat these salient channels equally, and they cannot fully exploit the useful information contained in salient channels to strengthen the common information of intra-class persons and the distinctive information of inter-class persons simultaneously. To this end, we introduce a multi-task strategy [15], [16] to construct a multi-stream refining network to learn discriminative features more effectively from hierarchical subnets, in which classification and measurement tasks are carried out for multi-stage refined features.

This paper has three major contributions:

(1) We design a multi-stream refining model to gradually enhance the focus on salient channels from multiple stages and fully exploit their available discriminant information in deep feature space.

(2) A hierarchical multi-task learning framework is proposed to learn more discriminant features from a multi-stream structure, which can systematically integrate the advantages of multiple loss tasks by sharing the information of lower stage subnets step by step.

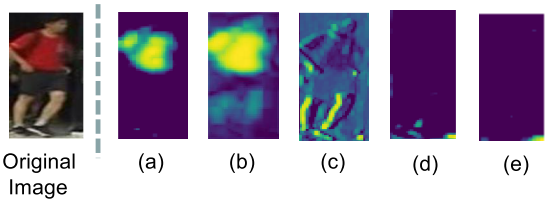

FIGURE 1. The left side of the dotted line is the original image in the Market-1501 dataset, and the right images (a), (b), (c), (d), (e) are the results of the visualization of some of the channels after the first pooling layer. 1) (a) is the result of the visualization of the 8-th channel, which only contains the salient information of the top red area of the original image. 2) (b) is the visualization of the 36th channel, which expresses the information of the entire body area of a person; 3) (c) is the visualization of the 6-th channel, and the salient channel for expressing the texture information is obtained, which is superior to the color information for the case where the illumination change is strong. 4) (d) and (e) are the results of the 23-rd and 32nd channels. Given low response channels like these, they can reduce the interference of channels that focus only on the background area on person re-ID.

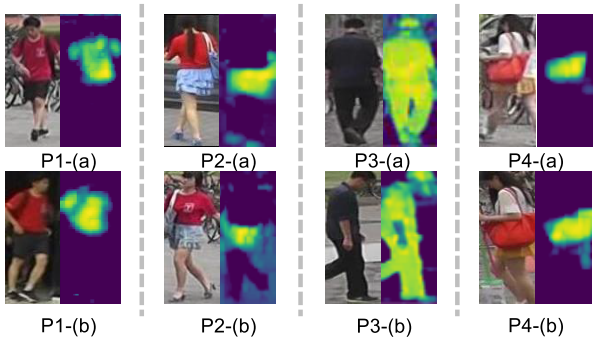

FIGURE 2. P1, P2, P3, and P4 represent four persons of different identities, (a) and (b) represent images taken by the person under different camera viewpoints. The left side color images in each column are all from the Market-1501 dataset, and the corresponding right side images are the result of visualizing some of the network channels.

(3) Our method is validated on three benchmark datasets, including the DukeMTMC-reID, Market1501, and MSMT17. Extensive ablation experiments demonstrate the effectiveness of each module in our proposed approach.

The rest of the paper is organized as follows. Section II reviews related works. In section III, details of our model structure are described. Section IV describes the experimental settings and discusses the validity of our model. Section V is our conclusion.

II. RELATED WORK

The whole process of person re-ID [10], [17] is mainly divided into two phases: feature representation and similarity measurement.

A. FEATURE REPRESENTATION

A large number of early traditional methods [3]–[5], [18]–[20] were applied to person re-ID to extract discriminative features. In these methods, Farenzena *et al.* [18] divided the head, trunk, legs, and left-right symmetrical axis of the human body through preprocessing, extracted the color and texture features of other areas besides the head, and weighted the features for person re-ID. Cheng *et al.* [19] used the color

features of each body part to accurately match. Liao *et al.* [20] proposed the feature representation method Local Maximal Occurrence (LOMO). The literature [3]–[5] make full use of the features in salient regions. Zhao *et al.* [3] used salient patch matching to deal with spatial misalignment problems. An unsupervised saliency learning method is proposed to learn saliency measures for person re-ID. Martinel [4] *et al.* proposed a kernelized graph-based approach to detect the salient regions of a person, which was later used as a weighting tool in the feature extraction process.

With the development of deep learning and the rapid increase of re-ID data, methods based on deep feature representation have achieved remarkable success [8]–[10], [21]–[28]. In these approaches, Zheng *et al.* [8], Zhong *et al.* [9], Qian *et al.* [29], Wei *et al.* [21] respectively applied DCGAN [30], CycleGAN [31], PN-GAN [29], PTGAN [21] to generate person images to make the model better optimized by expanding the dataset. To alleviate the influence of occlusion in person re-ID, Miao *et al.* [32] designed a model and checked the occluded area to generate corresponding features. Finally, this information is excluded during the matching process. Li *et al.* [33] proposed a joint learning framework and multi-scale technology to deal with the problem of low information content at low resolution. Sarfraz *et al.* [23] incorporated pose information into person re-ID and proposed a PSE network. Wei *et al.* [27] proposed the Global-Local-Alignment Descriptor (GLAD), which used human body key points to divide the image into three parts: head, upper body, and lower body. The three parts and the whole image are sent into a convolutional neural network shared by parameters to extract local and global features. Zhao *et al.* [10] extracted the features of seven different body regions of a person and fused them to obtain a better feature representation. Li *et al.* [22] believed that the problems of misalignment and background clutter have an impact on person re-ID, so the HA-CNN network was proposed. He *et al.* [26] mainly focused on the problem of occlusion, extracted the feature map through full convolution network, and then analyzed the similarity by deep spatial feature reconstruction. Tian *et al.* [28] proved the influence of background to person re-ID through experiments, and designed the network structure to extract the global features from the whole image and the local features from the head, upper body, and lower body after person segmentation. The global and local features were combined for person re-ID. Song *et al.* [25] also focused on the impact of the background, they designed a comparative attention model based on segmented person contour to learn person features independent of the background. Nevertheless, as far as we know, in person re-ID, most methods based on deep feature representations ignored that different channels have different effects. Most methods only treated them equally and did not highlight the importance of salient channels for person re-ID.

We are inspired by the great success of the deep learning network in the field of computer vision and pattern recognition [34]–[37]. In a deep network, different channels represent different salient information, the more salient channels are, the more helpful to identify a person. In our model, a self-supervised channel significance learning module is proposed to learn different weights for different channels. Salient channels will get higher weights.

B. SIMILARITY MEASUREMENT

At present, a large number of similarity measure methods are used in person re-ID [38]–[43]. Kostinger *et al.* [43] proposed KISSME's metric learning method. Liao [20] *et al.* used a quadratic linear discriminant analysis method to obtain a metric function for calculating the similarity of samples from different perspectives. Wang *et al.* [39] used a triplet training sample and a triple loss function to measure the similarity between images. Ding *et al.* [40] applied triplet loss to person re-ID for the first time. Cheng *et al.* [38] believed that the original triplet loss function only considers the inter-class distances between positive and negative samples while ignoring the intra-class distances between positive samples, so an improved triplet loss function was proposed. Chen *et al.* [11] also improved the triplet loss function by proposing a quadruplet loss function. It requires four input images at a time, including two positive samples and two negative samples from different individuals. The quadruplet loss function takes into account the absolute distances between positive and negative samples.

Many network models [44]–[49] considered person re-ID as a classification problem. Generally speaking, the softmax function is used to calculate the probability belonging to each class, and the cross-entropy loss function is adopted. In our network, a hierarchical multi-task learning framework is designed to optimize the multi-stream refining model to learn more discriminative features.

III. MODEL DESIGN

A. NETWORK ARCHITECTURE

Currently, most of the methods [8], [11], [13] adopt Zheng's model [56][57] as backbones which is based on ResNet-50 [35]. The ResNet-50 residual network contains a large number of residual blocks [35]. They are implemented with shortcut connections, and the input and output features are overlapped element-wise. Based on the advantages of the residual network and in order to illustrate the performance of the proposed network, we choose the ResNet-50 network as the backbone and divide it into five stages (subnets). The overall framework of the model is shown in Fig. 3. In the first stage of the network, it contains a convolution and pooling operation to the input image. After the first stage, each stage contains some residual blocks [35]. The output channel size of each stage is half of before, and the number of channels is twice as many as before. At the end of the backbone, the last fully connected layer has 1000 neurons. In addition, we have made the following changes.

(1) In the proposed network, we add a channel significance self-learning function as a refining feature sub-module

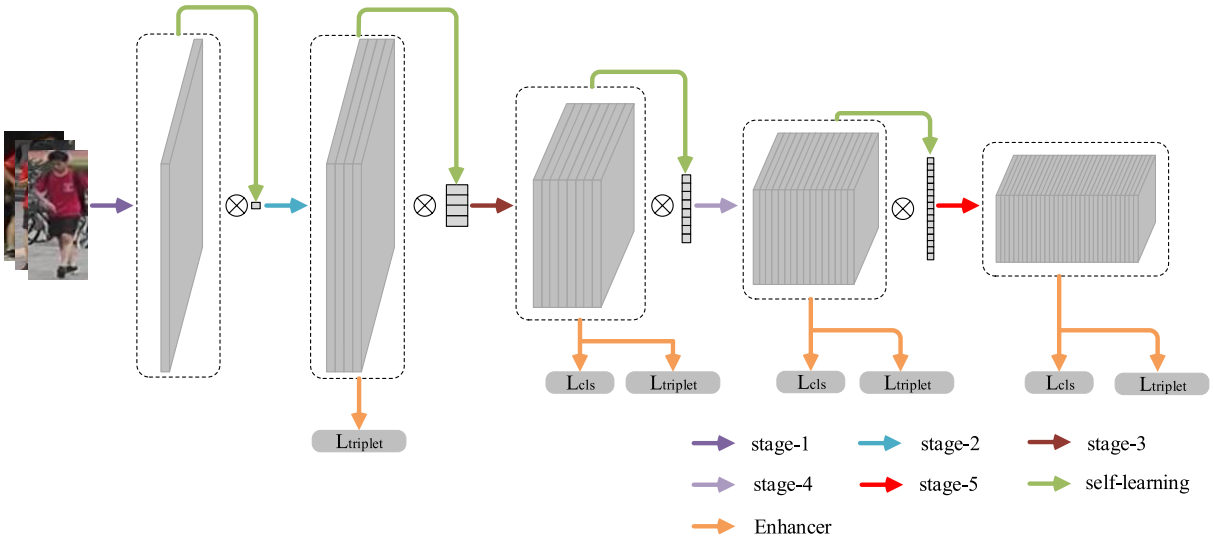

FIGURE 3. Our multi-stream refining network (MRNet).

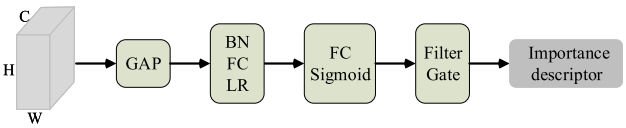

FIGURE 4. Our channel significance self-learning sub-module. GAP is the global average pooling, BN is the batch normalization, LR is the LeakyReLU activation function, FC is the full connection layer, Sigmoid is the sigmoid activation function, and Filter Gate represents the filtering of importance descriptors.

into the first four stages, respectively. This sub-module can learn the importance of descriptors of each channel. Their ranges are controlled from 0.3 to 1. The learned importance descriptors are multiplied with the corresponding channels to perform weighting operation. With the refining sub-module, the network can focus more on salient channels, so that the interference of the channel corresponding to the background clutter can be eliminated partially and the robustness of the re-ID model could be improved.

(2) We introduce an enhancement function as another refining feature sub-module to the network and add it into the last four stages, respectively. So our model becomes a multi-stream structure. The enhancement sub-module aims to learn the common intra-class information and distinctive inter-class information.

(3) Based on the above multi-stream structure, we devise a hierarchical multi-task learning framework. For the last three streams, we calculate multi-task losses, including classification and measurement tasks. Considering that the first stream is composed of the low-layer network, we only calculate the measurement loss. These changes can make the whole network form a multi-stream multi-task learning framework. Thus, multiple subnetworks can fully share information with each other in the learning process, which helps to learn more discriminative features and improve the generalization ability of the re-ID model.

B. MULTI-STREAM REFINING MODEL

Our multi-stream refining model consists of two sub-modules, i.e., channel significance self-learning and enhancement sub-module.

1) CHANNEL SIGNIFICANCE SELF-LEARNING SUB-MODULE

Using the channel significance self-learning sub-module, the weight of the channel can be adjusted continuously by reducing the loss iteratively in the process of optimizing the model. The sub-module structure is shown in Fig. 4.

Importance Descriptors. Suppose the feature learned in the i-th stage is $\mathbf{X}^{(i)} \in \mathbb{R}^{W\times H\times C}$, where $i \in \{1, 2, 3, 4, 5\}$ represents the five stages, $\mathbf{X}^{(i)} = [\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \cdots, \mathbf{x}_c^{(i)}]$, $\mathbf{x}_c^{(i)} \in \mathbb{R}^{W\times H}$, and $\mathbf{x}_c^{(i)}$ represents the c-th channel in the i-th stage. In the first four stages, the number of channels are 64, 256, 512, 1024 respectively. Furthermore, we can obtain the numerical descriptors $y_c^{(i)}$ corresponding to each channel through global average pooling operation (GAP) and express its importance.

$$y_c^{(i)} = \frac{1}{W \times H} \sum_{u=1}^{W} \sum_{v=1}^{H} \mathbf{x}_c^{(i)}(u, v) \tag{1}$$

where $\mathbf{x}_c^{(i)}(u, v)$ is the value of the c-th channel at (u, v) in the i-th stage. Then the numerical descriptor of the i-th stage can be written as $\mathbf{Y}^{(i)} = [y_1^{(i)}, y_2^{(i)}, \cdots, y_c^{(i)}]$.

After that, we use the sigmoid function to convert them into importance descriptors $\mathbf{Z}^{(i)}$, where the numerical descriptors are needed to be normalized and the output is integrated by two fully connected layers. Thus, the value range of the importance descriptor $\mathbf{Z}^{(i)} = [z_1^{(i)}, z_2^{(i)}, \cdots, z_c^{(i)}]$ is between 0 and 1., where $z_c^{(i)}$ represents the importance descriptor of the c-th channel in the i-th stage. $\mathbf{Z}^{(i)}$ is calculated by the following formula

$$\mathbf{Z}^{(i)} = \sigma(\Psi(\mathbf{Y}^{(i)})) \tag{2}$$

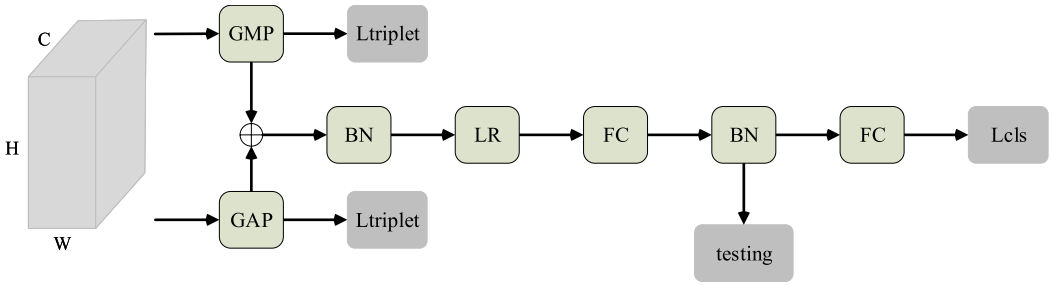

FIGURE 5. Our enhancement sub-module. GAP is the global average pooling, GMP is the global maximum pooling. BN is the batch normalization, LR is the LeakyReLU activation function, and FC is the full connection layer.

where Ψ denotes two fully connected layers, and σ is the sigmoid function. The number of neurons in the first fully connected layers is 512. The number of neurons in the second layer is the number of channels at this stage. Since LeakyReLU [50], [51] does not have zero-slope parts which can fix the "dying ReLU" problem, we select it as the activation function.

Filter Gate. In the experiment of channel visualization, we found that some channels mainly corresponding to the background area also carried a small amount of personal information. If these channels are assigned smaller weights, some personal information might be ignored. Therefore, we design a filter gate to control the importance of descriptor between 0.3 and 1. The new descriptor is defined as $\widetilde{\mathbf{Z}}^{(i)} = [\widetilde{z}_1^{(i)}, \widetilde{z}_2^{(i)}, \cdots, \widetilde{z}_c^{(i)}]$, where

$$\widetilde{z}_c^{(i)} = \begin{cases} 0.3 & z_c^{(i)} < 0.3 \\ z_c^{(i)} & 0.3 \le z_c^{(i)} < 0.9 \\ 1 & z_c^{(i)} \ge 0.9 \end{cases} \tag{3}$$

In the above formula, if the importance descriptor of a channel is lower than 0.3, its value will be reset to 0.3. If it is higher than 0.9, the value will be reset to 1.

Application of Importance Descriptor. The function of the importance descriptor is to weight the channels at each stage, so as to get the salient channel. The weighted channel $\widetilde{\mathbf{x}}_c^{(i)}$ can be expressed as follows

$$\widetilde{\mathbf{x}}_c^{(i)}(u, v) = \mathbf{x}_c^{(i)}(u, v) \times z_c^{(i)} \tag{4}$$

where $\mathbf{x}_c^{(i)}(u, v)$ is the value at (u, v) of the c-th channel in the i-th stage, $u \in \{1, 2, \cdots, W\}$, and $v \in \{1, 2, \cdots, H\}$. $\widetilde{\mathbf{x}}_c^{(i)}(u, v)$ is the weighted value at (u, v) of the c-th channel in the i-th stage. At last, the refined features $\widetilde{\mathbf{X}}^{(i)} = [\widetilde{\mathbf{x}}_1^{(i)}, \widetilde{\mathbf{x}}_2^{(i)}, \cdots, \widetilde{\mathbf{x}}_c^{(i)}]$ of the i-th stage will flow into the next stage.

2) ENHANCEMENT SUB-MODULE

In order to make full use of the features learned in each stage and strengthen the contribution of common and distinctive information in salient channels to the model loss, we construct an enhancement sub-module by combining global average pooling (GAP) and global maximum pooling (GMP). The sub-module structure is shown in Fig. 5. GAP can gather the average value of the spatial information in channels as the common feature information of the same kind of person. GMP can find the most salient part in channels as the distinctive feature information of different classes. Therefore, they help our model to calculate the intra-class similarity and inter-class dissimilarity.

The feature learned in the i-th stage is $\mathbf{X}^{(i)} \in \mathbb{R}^{W \times H \times C}$, where $i \in \{2, 3, 4, 5\}$ indicates the last four stages. We use GAP and GMP to obtain the metric features $\mathbf{Y}^{(i)}$ and $\widetilde{\mathbf{Y}}^{(i)}$, respectively.

$$\mathbf{Y}^{(i)} = G_1(\mathbf{X}^{(i)}) \tag{5}$$

$$\widetilde{\mathbf{Y}}^{(i)} = G_2(\mathbf{X}^{(i)}) \tag{6}$$

where G_1 and G_2 represent GAP and GMP, respectively. Through aggregation and a full connection layer, we get the refined features $\mathbf{M}^{(i)}$.

$$\mathbf{M}^{(i)} = f_1(\mathbf{Y}^{(i)} \oplus \widetilde{\mathbf{Y}}^{(i)}) \tag{7}$$

where $\oplus$ represents the element-wise addition at the corresponding position of the features, f_1 is a fully connected layer, and the number of neurons is 512. Then $\mathbf{M}^{(i)}$ is fed into another full connection layer to get classification feature $\mathbf{N}^{(i)}$ as follows.

$$\mathbf{N}^{(i)} = f_2(\mathbf{M}^{(i)}) \tag{8}$$

where f_2 is a fully connected layer, and the number of neurons is the number of classes in the training set. $\mathbf{M}^{(i)}$ will also be used in the subsequent testing experiment.

We introduce the enhancement sub-module into the last four stages respectively to constitute a multi-stream structure. In testing, we fuse the features $\mathbf{M}^{(3)}$, $\mathbf{M}^{(4)}$, $\mathbf{M}^{(5)}$ of the last three streams as follows.

$$\mathbf{M} = \phi_{concat}(\mathbf{M}^{(3)}, \mathbf{M}^{(4)}, \mathbf{M}^{(5)}) \tag{9}$$

where ϕ_{concat} is a concatenation operation. The dimension of $\mathbf{M}$ is 1536.

C. HIERARCHICAL MULTI-TASK LEARNING FRAMEWORK

Different layers of the network may contain discriminative information of the person. In order to fully refine the useful information, we design a hierarchical multi-task learning framework based on the above multi-stream structure to form the MRNet. As shown in Fig.3, in addition to the first stream

with only one measurement task, the other three streams contain classification and measurement tasks simultaneously.

The objective function L of our model includes two task losses.

$$L = \lambda_1 \cdot L_{tri} + \lambda_2 \cdot L_{cls} \tag{10}$$

where L_{tri}, L_{cls} are the losses caused by a classification task and measurement task respectively, and λ_1, λ_2 are the weight parameters balancing two tasks.

1) MEASUREMENT TASK

In order to reduce the intra-class distance between the same persons and increase the inter-class distance between different persons, we adopt a triplet loss function to calculate the measurement loss L_{tri} in four streams. In addition, each stream consists of local measurement loss and global measurement loss. The whole measurement task can be represented as follows.

$$L_{tri} = \frac{1}{4} \cdot (L_{tri}^{(1)} + L_{tri}^{(2)} + L_{tri}^{(3)} + L_{tri}^{(4)}) \tag{11}$$

where $L_{tri}^{(1)}, L_{tri}^{(2)}, L_{tri}^{(3)}, L_{tri}^{(4)}$ denote the measurement losses in the four streams respectively.

$$L_{tri}^{(i)} = \frac{1}{2} \cdot (L_{tri}^{(i-GAP)} + L_{tri}^{(i-GMP)}) \tag{12}$$

where $i \in \{1, 2, 3, 4\}$, i indicates the i-th stream, $L_{tri}^{(i-GAP)}$ is global measurement loss, and $L_{tri}^{(i-GMP)}$ is local measurement loss. They are expressed as follows.

$$L_{tri}^{(i-GAP)} = [\mathbf{Y}_p^{(i+1)} + \mathbf{Y}_n^{(i+1)} + \alpha]_+ \tag{13}$$

$$L_{tri}^{(i-GMP)} = [\widetilde{\mathbf{Y}}_p^{(i+1)} + \widetilde{\mathbf{Y}}_n^{(i+1)} + \alpha]_+ \tag{14}$$

After GAP and GMP, we can get $\mathbf{Y}^{(i+1)}$ and $\widetilde{\mathbf{Y}}^{(i+1)}$ in the i-th stream. $\mathbf{Y}^{(i+1)}$ indicates the global features of pedestrians. $\widetilde{\mathbf{Y}}^{(i+1)}$ stands for the local feature of the most salient region. $\mathbf{Y}_p^{(i+1)}$ and $\mathbf{Y}_n^{(i+1)}$ ($\widetilde{\mathbf{Y}}_p^{(i+1)}$ and $\widetilde{\mathbf{Y}}_n^{(i+1)}$) are the distance between positive and negative sample pairs respectively. α is the distance parameter of triplet loss function. $[x]_+$ is $max(0, x)$. In our experiments, α is set to 0.3.

2) CLASSIFICATION TASK

We calculate the weighted sum L_{cls} of the metric features $\mathbf{N}^{(3)}$, $\mathbf{N}^{(4)}$ and $\mathbf{N}^{(5)}$ in the 2nd, 3rd and 4-th stream as the classification loss.

$$L_{cls} = L_{(\mathbf{N}^{(3);label})} + L_{(\mathbf{N}^{(4);label})} + L_{(\mathbf{N}^{(5);label})} \tag{15}$$

where $L_{(\mathbf{N}^{(3);label})}$, $L_{(\mathbf{N}^{(4);label})}$, $L_{(\mathbf{N}^{(5);label})}$ represent the classification loss caused by softmax function in the last three streams respectively.

$$L_{(\mathbf{N}^{(i+1);label})} = -\log\left[\frac{\exp(\mathbf{N}_{label}^{(i+1)})}{\sum_i^{class}\exp(\mathbf{N}_j^{(i+1)})}\right] \tag{16}$$

where $\mathbf{N}_j^{(i+1)}$ is the predicted score of the j-th person in the i-th stream. $\mathbf{N}_{label}^{(i+1)}$ denotes the predicted score of the person *label* in the i-th stream. $i \in \{2, 3, 4\}$ indicates the 2nd, 3rd, 4-th stream respectively.

TABLE 1. The person re-ID datasets used in our experiment and their details.

Dataset	Market-1501	DukeMTMC-reID	MSMT17
Indentities	1,501	1,404	4,101
BBoxes	32,668	36411	126,441
Cameras	6	8	15
Label method	DPM	manually labeled	Faster RCNN
Train # imgs	12,936	16,522	32,621
Train # ids	751	702	1,041
Test # imgs	19,732	19,889	93,820
Test # ids	750	702	3,060

IV. EXPERIMENT

We evaluate the proposed network model on three large scale person re-ID datasets and compare the proposed method with the state-of-the-art.

A. DATASETS AND SETTINGS

Table 1 describes the three benchmark datasets we used.

Market-1501 dataset [52]. The Market-1501 dataset is one of the most widely used benchmark datasets in person re-ID. The dataset consists of 1501 persons captured by 6 cameras (5 high-definition cameras and 1 low-definition camera) and 32,668 bounding boxes detected by the DPM detector. Each of these people has appeared on at least two cameras. The training set has 751 people and contains 12,936 images. The test set has 750 people and contains 19,732 images.

MSMT17 dataset [21]. The MSMT17 is a large, challenging dataset. It was released in 2018. The dataset is captured by 15 cameras (12 outdoor cameras and 3 indoor cameras) at different time periods, which includes 4,101 persons and 126,441 detected bounding boxes. The training set has 1041 persons and 32,621 bounding boxes. The test set consists of 3,060 persons and 93,820 bounding boxes. The dataset covers multiple viewpoints and time periods. Compared with the earlier datasets, it fully considers the impact of complex viewpoints and significant illumination changes in person re-ID.

DukeMTMC-reID dataset [8]. DukeMTMC-reID dataset is captured by 8 cameras. There are both 702 persons in the training set and test set. The training set has 16,522 images. In the test set, the query has 2,228 images and gallery has 17,661 of them.

Evaluation metrics. Cumulative matching characteristics (CMC) Rank-1 accuracy and mAP are the most commonly used evaluation metrics. Now it reaches an agreement to utilize Rank-1 and mAP for person Re-ID task. Therefore, we also take Rank-1 and mAP as evaluation metrics.

B. IMPLEMENTATION DETAILS

1) SYSTEM SETTINGS

The proposed MRNet is implemented with Pytorch deep learning framework, including torch 1.0.1, CUDA 8.0.61, cudnn 7.1.2. The python version is 3.6.8. The hardware of server contains 12G GeForce RTX 1080Ti, Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz. The operating system is Ubuntu 16.04.6 LTS.

TABLE 2. Rank-1 and mAP accuracies of three variants on the Market-1501 dataset. RR means re-ranking.

Variant	Rank-1	mAP	Rank-1(RR)	mAP(RR)
baseline	93.08%	81.44%	94.77%	92.88%
MRNet(without MM)	95.48%	87.48%	96.05%	94.37%
MRNet	95.67%	88.51%	96.11%	94.64%

FIGURE 6. Different identities have some different salient information. When we identify a person, we will judge them more according to the salient information they have. These images are from the Market-1501 dataset.

2) TRAINING SETTINGS

Cross entropy loss with label smoothing [53] is used for training. The parameters of MRNet are initialized from ImageNet pretrained weights. Euclidean distance is utilized for a person matching. Weight decay is set to 0.0005 and momentum is set to 0.9. The batch size is set to 32, 64, 64 for Market-1501, MSMT17 and DukeMTMC-reID datasets respectively. The stochastic gradient descent (SGD) algorithm is used to train the network for 500 epochs. The learning rate starts from 0.002 and is decayed by 0.1 every 10 epochs between 60 and 130 epochs. The values of the weight parameters λ_1, λ_2 in Equation (10) are set to 1. Data augmentation includes random flip and random erasing [54]. Images are resized to 288×144. In order to improve the performance of the model, we apply the re-ranking method [17]. Re-ranking combines the original distance and the Jaccard distance to produce a new ranking result, which has a better improvement for the final performance.

C. ABLATION STUDY

In order to prove the effectiveness and relevance of each improvement, we propose the following three variants, each of them is tested on the Market-1501 dataset, and the results are shown in Table 2.

Variant 1 (denote baseline). The backbone based on ResNet-50 is modified by Zheng *et al.* [13], [55]. We add the preprocessing operations mentioned above during the experiments. And use it as our baseline in this paper.

Variant 2 (denote MRNet(without MM)). Our multi-stream refined feature network.

Variant 3 (denote MRNet). Based on MRNet, we propose a multi-stream multi-task learning strategy to optimize the model.

1) THE IMPACT OF CHANNEL SIGNIFICANCE SELF-LEARNING SUB-MODULE

In the case that the face is not very clear when people identify whether the 2 persons are the same one, they often rely on some salient information of the person. For example, in Fig. 6, person P1 has the salient information of a red top and black shorts. The salient information of P2 is a red shirt and a blue skirt. The P3 salient information is the whole black suit and P4 is the red bag. Normally it is easier to judge a person's identity based on this salient information. This salient information will not change as the viewpoint changes. In the deep convolutional network, different channels of the same layer express different salient information of the person. As illustrated in Fig. 7, we visualize partial channels of four persons and the importance descriptors that they learned through the module. In P1, the salient channels of red short sleeves and black shorts get higher weights with our model. The salient channels of the blue short skirt and the red top in P2 learn higher weights. The salient channels of the whole black suit expressed in P3 get higher weights. In person P4, the salient channels of the red bag get higher weights. For channels such as P1-(c), P2-(d), P3-(c), and P4-(d), they have lower weights learned by our model. The interference of viewpoints and background area is reduced. As shown in Table 2, by adding a channel significance self-learning module, the accuracies of Rank-1 and mAP are improved by 1.3% and 2.2% respectively on the Market-1501 dataset compared to the baseline.

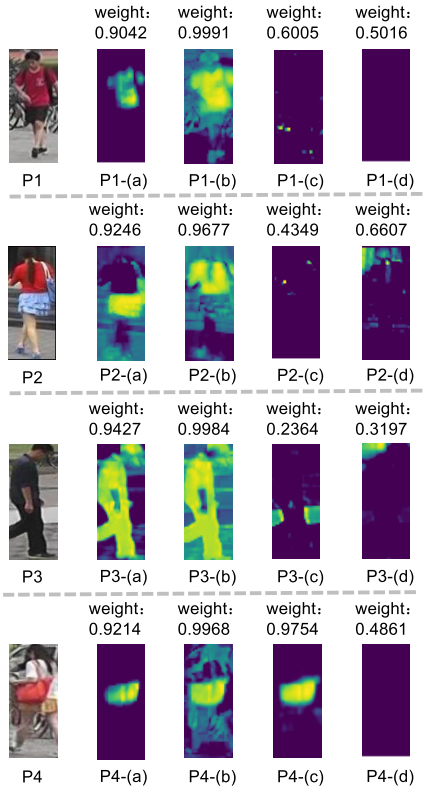

FIGURE 7. Different channels express different saliency information of a person. We visualize the channels of person P1, P2, P3, and P4, as shown in (a), (b), (c), and (d). With our improvement, the model learns different weight parameters for different salient channels, which makes the features with salient information gets higher weights and reduces the interference of channels expressing the background information. Thereby improving the discriminative power of the model.

TABLE 3. The effect of setting different minimum weights in channel significance self-learning module in our model MRNet. The experiments were conducted on Market-1501 dataset.

Method	Rank-1	mAP	Rank-1(RR)	mAP(RR)
MRNet(0)	95.39%	88.05%	96.08%	94.38%
MRNet(0.2)	95.16%	88.30%	96.17%	94.62%
MRNet(0.3)	95.67%	88.51%	96.11%	94.64%
MRNet(0.4)	95.28%	88.32%	95.99%	94.38%

TABLE 4. The impact of global average pooling (GAP) and global maximum pooling (GMP) on our model MRNet. The experiments were tested on Market-1501 dataset.

Method	Rank-1	mAP	Rank-1(RR)	mAP(RR)
MRNet(only GAP)	95.07%	88.37%	95.90%	94.40%
MRNet(only GMP)	95.37%	87.81%	95.75%	94.39%
MRNet	95.67%	88.51%	96.11%	94.64%

In the channel significance self-learning module, we set a filter gate to constrain the importance descriptors. In the process of visualizing the channels, we found that some channels with lower weights still express a small amount of a person's information. If they obtain lower weights during the learning process, the model will ignore this part of the information. Therefore, we constrain the minimum values of the importance descriptors and limit the minimum weights of the importance descriptors to 0, 0.2, 0.3, and 0.4 in our experiments. The experimental results are shown in Table 3. When the minimum weights of the importance descriptors are controlled at 0.3, the model achieves the best recognition performance. Compared with MRNet(0), the model improved by 0.5 on mAP.

2) THE IMPACT OF ENHANCEMENT SUB-MODULE

As shown in Table 2, compared with baseline, the accuracies of rank-1 and mAP increased by 1.5% and 4.4%, respectively. Then we test the features of each stream in MRNet separately, and the comparison results are shown in Fig. 8. MRNet (stream2), MRNet (stream3), MRNet (stream4) represent the results of testing using only the features of the 2nd, 3rd, 4-th stream, respectively. MRNet(stream 3,4) represents the result of fusing the features of the 3rd and 4-th streams during the test phase. MRNet represents the result of fusing the features of the 2nd, 3rd, and 4-th streams during the test phase. From the figure, we can intuitively see that the best performance is achieved by fusing multi-stream features, and rank-1 and mAP have achieved better improvement. It makes full use of the multi-stream features.

The enhancement module of each stream contains two parts: global average pooling (GAP) and global maximum pooling (GMP), and we perform experiments on models that only include GMP and GAP. The experimental results are shown in Table 4. We can intuitively see that when MRNet only contains a GAP, the performance on the mAP is better. GAP calculates average information from different locations in the channels, so that the model has a global use of the information on the channel. When MRNet only contains GMP, the model pays more attention to the more salient

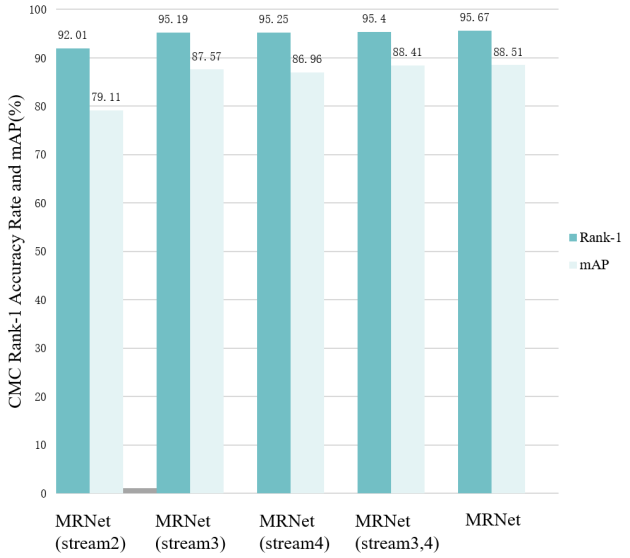

FIGURE 8. MRNet (stream2), MRNet (stream3), MRNet (stream4) represent the results of using only the features of the 2nd, 3rd, 4-th stream, respectively. MRNet(stream3, 4) represents the result of fusing the features of the 3rd and 4-th streams to test. MRNet represents the result of fusing the features of the 2nd, 3rd and 4-th streams during the test phase. All the above models are tested on Market-1501 dataset.

TABLE 5. The impact of classification task and measurement task. The experiments were tested on Market-1501 dataset.

Method	Rank-1	mAP	Rank-1(RR)	mAP(RR)
MRNet(only measurement task)	90.64%	77.98%	91.95%	88.02%
MRNet(only classification task)	95.48%	87.48%	96.05%	94.37%
MRNet	95.67%	88.51%	96.11%	94.64%

local areas in the channel, and the model has higher accuracy in Rank-1. Through the enhancement module, we have a holistic consideration of each pixel in each channel and take advantage of the most salient area in channels to identify persons. The accuracy of Rank-1 and mAP have improved significantly.

3) THE IMPACT OF MULTI-STREAM MULTI-TASK LEARNING FRAMEWORK

As shown in Table 5, MRNet (only measurement task) represents the result of optimizing the model only using multi-stream measurement task, and MRNet (only classification task) represents the result of optimizing the model only using multi-stream classification tasks. The convergence curves for different task losses are shown in Figure 9, where the classification loss is larger due to the use of label smoothing. In the multi-stream measurement task, the triplet loss can pull the distance between the images of the same person, and image pairs with different people are pushed away at the same time. The design of the multi-stream classification task can learn the characteristics of a person in the same category in a more accurate way. The multi-stream multi-task learning framework can better optimize the model and extract more robust features.

TABLE 6. Comparison of our approach with the published state-of-the-art on the Market-1501 dataset.

Method		Rank-1	mAP
SpindleNet [11]	CVPR2017	76.90%	-
Cross-view [58]	TIP2018	80.31%	59.68%
SVDNet [13]	ICCV2017	82.30%	62.10%
LSRO [8]	CVPR2017	83.97%	66.07%
PDC [59]	ICCV2017	84.14%	63.41%
FMN [60]	arXiv2017	85.99%	67.12%
PAN [61]	TCSVT2019	86.67%	69.33%
PIE [62]	TIP2019	89.06%	70.69%
PNGAN [29]	ECCV2018	89.43%	72.58%
CamStyle [9]	CVPR2018	89.49%	71.55%
BISAA [63]	TCSVT2019	91.2%	75.1%
HA-CNN [22]	CVPR2018	91.20%	75.70%
PCB(DropEasy2d) [64]	Access2019	93.80%	78.30%
PCB [56]	ECCV2018	93.80%	81.60%
AANet [65]	CVPR2019	93.93%	83.41%
HPDN [66]	TETCI2020	94.00%	81.20%
CASN [67]	CVPR2019	94.4%	82.8%
IANet [68]	CVPR2019	94.4%	83.1%
OSNet [14]	ICCV2019	94.8%	84.9%
DG-Net [55]	CVPR2019	94.8%	86.0%
CRAN [69]	TCSVT2019	94.9%	84.9%
FPR [70]	ICCV2019	95.42%	86.58%
BNNeck+RR [71]	IVC2020	95.5%	93.2%
MGN [57]	ACM MM2018	95.7%	86.9%
pyramidal [72]	CVPR2019	95.7%	88.2%
FAN [73]	ICCV2019	**96.1%**	84.7%
SCSN(4 stages) [74]	CVPR2020	95.70%	88.50%
HOReID [75]	CVPR2020	94.20	84.90 %
MRNet		95.67%	**88.51%**
MRNet + RR		96.11%	94.64%

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

We compare some state-of-the-art methods in this section. In order to make a fair comparison with other methods, we compare against the accuracies mentioned in their corresponding papers.

1) RESULTS ON MARKET-1501 DATASET

The Market-1501 dataset is the most widely used dataset for person re-ID. The dataset images were taken by a low-resolution camera and 5 resolution cameras. It takes into account the interference of the resolution on person re-ID. There are lots of recent works on this dataset, numerical results of our MRNet and other state-of-the-art methods are shown in Table 6.

SpindleNet [11] utilizes body structure information for person re-ID. It extracts the body region features and fuses these features. Compared with SpindleNet, the accuracy of Rank-1 of our model has improved by 18.77%. LSRO [8] applies GAN to generate images to expand datasets. The backbone network is also based on Resnet-50. Our proposed model outperforms LSRO for 11.7% and 22.44% in terms of Rank-1 accuracy and mAP respectively. CamStyle [9] uses CycleGAN to generate images of different camera styles. Our approach improves CamStyle by 6.18% and 17% in terms of the accuracy of Rank-1 and mAP respectively. PCB [56] trains six classifiers in the training phase by training on image blocks and proposes an RPP model to further improve the discrimination of features. Compared with PCB, our Rank-1 accuracy and mAP has increased by 1.8% and 6.9%, respectively. MGN [57] synthetically utilizes the global and fine-grained features in images, and then learns the model through multiple branches. It uses the features from multiple branches for testing. Our Rank-1 is very close to MGN, but mAP is 1.61% higher than MGN. Compared with the above methods, our MRNet model has achieved very competitive results. Rank-1 and mAP reached 95.67% and 88.51%, respectively. After re-ranking, Rank-1 and mAP reached 96.11% and 94.64%.

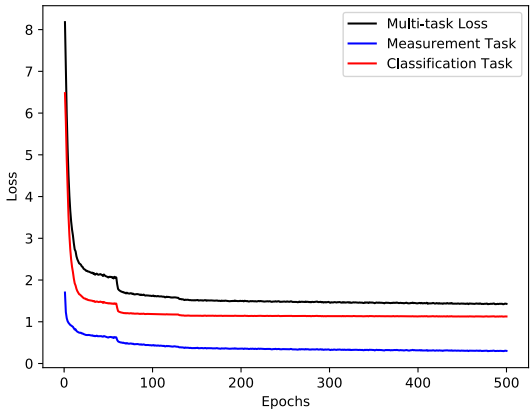

FIGURE 9. Loss curves of different tasks during the training process.

TABLE 7. Comparison of our approach with the published state-of-the-art on MSMT17 dataset.

Method		Rank-1	mAP
GoogLeNet [36]	CVPR2015	47.60%	23.00%
PDC [59]	ICCV2017	58.00%	29.70%
GLAD [27]	ACM MM2017	61.40%	34.00%
PCB [56]	ECCV2018	68.2%	40.4%
BISAA [63]	TCSVT2019	68.7%	39.1%
IANet [68]	CVPR2019	75.5%	46.8%
DG-Net [55]	CVPR2019	77.2%	52.3%
CRAN [69]	TCSVT2019	78.7%	52.4%
OSNet [14]	ICCV2019	78.7%	52.9%
SCSN(4 stages) [74]	CVPR2020	83.80%	58.50%
RGA-SC [76]	CVPR2020	80.30%	57.50%
MRNet		**87.53%**	**65.54%**

2) RESULTS ON MSMT17 DATASET

To the best of our knowledge, the MSMT17 is the latest and largest public dataset at present, in which data is obtained from multiple viewpoints at different periods. It takes the interference of illumination, viewpoints, and other issues fully into account. Due to the difficulty of data collection, the authors prefer to explore more effective and robust training strategies and models, the ratio of the training set and test set is set to 1:3. The numerical results of our MRNet and other state-of-the-art methods are shown in Table 7. Compared with the CRAN [69], our Rank-1 accuracy and mAP has increased by 8.8% and 13.1%, respectively. OSNet [14] proposes a new lightweight CNN structure, which can learn the features of isomorphism and heterogeneous scale better. Compared with the OSNet, our Rank-1 accuracy and mAP has increased by 8.8% and 12.6%, respectively. The data in Table 7 is

TABLE 8. Comparison of our approach with the published state-of-the-art on DukeMTMC-roeID dataset.

Method		Rank-1	mAP
LSRO [8]	CVPR2017	67.68%	47.13%
SVDNet [13]	ICCV2017	76.70%	56.80%
PNGAN [29]	ECCV2018	73.58%	53.20%
CamStyle+RE [9]	CVPR2018	78.32%	57.61%
End-to-End Deep [77]	CVPR2018	80.3%	63.2%
HA-CNN [22]	CVPR2018	80.50%	63.80%
MLFN [44]	CVPR2018	81.00%	62.80%
DuATM [6]	CVPR2018	81.82%	64.58%
MHN-6(IDE) [78]	ICCV2019	87.5%	75.2%
HOReID [75]	CVPR2020	86.90 %	75.60%
M^3+ResNet50 [79]	CVPR2020	84.70 %	68.50%
MRNet		**89.32%**	**78.99%**
MRNet+RR		91.02%	89.47%

very intuitive. Compared with other methods in the MSMT17 dataset, our MRNet model has a very significant performance improvement. The accuracy of rank-1 is 87.53%, and mAP is 65.54%.

3) RESULTS ON DUKEMTMC-REID DATASET

A lot of methods are applied to DukeMTMC-reID dataset and have got the good achievement. We compared our method with them. The numerical comparison results are shown in Table 8.

SVDNet [13] is modified based on the resnet50 network. Eigen layer is added before the final fully connected layer to obtain more discriminative features. Compared with SVDNet, our Rank-1 accuracy has increased by 12.62% and mAP has increased by 22.19%. PNGAN [29] utilize pose-normalization GAN to generate images of different poses of the same person to expand the dataset. The model is also modified based on the Resnet50 network. The proposed method outperforms PNGAN by 15.74% and 25.79% with respect to Rank-1 accuracy and mAP. HA-CNN [22] combines a person's local and global information by learning the hard region-level features and soft pixel-level attention features. The model we proposed improves HA-CNN by 15.19% and 8.82% in terms of mAP and Rank-1 accuracy. MLFN [44] proposed a new multi-level factor analysis network to learn potential discrimination without manual labeling. HSCNet gains 16.19% and 8.32% in terms of the mAP and Rank-1 accuracy.

V. CONCLUSION

In this paper, a novel MRNet model is proposed for person re-ID. First, we introduce a multi-stream refining model to adaptively control the importances of different salient channels by a self-learning sub-module. Then an enhancement sub-module is adopted to further refine the common and distinctive information from these channels from the multi-stage subnets. Finally, a hierarchical multi-task framework is employed to fully learn the discriminative features from the hierarchical subnets. Extensive experiments on three public datasets validate the state-of-the-art performance of the proposed MRNet and ablation studies illustrate the effectiveness of our proposed method. However, the proposed MRNet is time consuming due to deep architecture. Besides, the proposed method is just a stage of person Re-ID system since the pedestrian images are given. As a result, when facing real-world data, the proposed model may become vulnerable. Therefore, a network with pruning strategies for pedestrian detection and person Re-ID will be explored for future work. In fact, existing person Re-ID datasets are built-in ideal condition. Hence, designing an end-to-end Re-ID model to confront massive video data in reality will be meaningful future work.

REFERENCES

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

[2] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5098–5107.

[3] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 356–370, Feb. 2017.

[4] N. Martinel, C. Micheloni, and G. L. Foresti, "Kernelized saliency-based person re-identification through multiple metric learning," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5645–5658, Dec. 2015.

[5] L. An, M. Kafai, S. Yang, and B. Bhanu, "Person reidentification with reference descriptor," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 4, pp. 776–787, Apr. 2016.

[6] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5363–5372.

[7] 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. (CVPR)*, Jun. 2018, pp. 8042–8051.

[8] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 3754–3762.

[9] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5157–5166.

[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, pp. 907–915.

[11] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 403–412.

[12] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 994–1003.

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

[14] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3702–3712.

[15] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang, "PAMTRI: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 211–220.

[16] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*. [Online]. Available: http://arxiv.org/abs/1706.05098

[17] 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)*, Jul. 2017, pp. 3652–3661.

[18] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2360–2367.
[19] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2011, vol. 1, no. 2, p. 6.
[20] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.
[21] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 79–88.
[22] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2285–2294.
[23] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 7, Jun. 2018, p. 8.
[24] 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. (CVPR)*, Jun. 2018, pp. 4099–4108.
[25] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1179–1188.
[26] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7073–7082.
[27] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-local-alignment descriptor for pedestrian retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 420–428.
[28] 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. (CVPR)*, Jun. 2018, pp. 5794–5803.
[29] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2018, pp. 661–678.
[30] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434
[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
[32] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 542–551.
[33] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3765–3773.
[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[38] 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, pp. 1335–1344.
[39] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1386–1393.
[40] 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.
[41] 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.
[42] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737
[43] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2288–2295.
[44] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2109–2118.
[45] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3908–3916.
[46] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. 22nd Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2014, pp. 34–39.
[47] 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. (CVPR)*, Jun. 2016, pp. 1288–1296.
[48] 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. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 135–153.
[49] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 791–808.
[50] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in *Proc. 22nd Int. Conf. Digit. Signal Process. (DSP)*, Aug. 2017, pp. 1–5.
[51] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*. [Online]. Available: http://arxiv.org/abs/1505.00853
[52] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2015, pp. 1116–1124.
[53] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[54] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv:1708.04896*. [Online]. Available: http://arxiv.org/abs/1708.04896
[55] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2138–2147.
[56] 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)*. Cham, Switzerland: Springer, 2018, pp. 501–518.
[57] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.
[58] A. Borgia, Y. Hua, E. Kodirov, and N. M. Robertson, "Cross-view discriminative feature learning for person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5338–5349, Nov. 2018.
[59] 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)*, Oct. 2017, pp. 3980–3989.
[60] G. Ding, S. Khan, Z. Tang, and F. Porikli, "Let features decide for themselves: Feature mask network for person re-identification," 2017, *arXiv:1711.07155*. [Online]. Available: http://arxiv.org/abs/1711.07155
[61] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2019.
[62] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.

[63] X. Liu, S. Bi, S. Fang, and A. Bouridane, "Bayesian inferred self-attentive aggregation for multi-shot person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3446–3458, Oct. 2020.
[64] H. Wang, T. Fang, Y. Fan, and W. Wu, "Person re-identification based on DropEasy method," *IEEE Access*, vol. 7, pp. 97021–97031, 2019.
[65] C.-P. Tay, S. Roy, and K.-H. Yap, "AANet: Attribute attention network for person re-identifications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7134–7143.
[66] Z. Zhang and M. Huang, "Person re-identification based on heterogeneous part-based deep network in camera networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 1, pp. 51–60, Feb. 2020.
[67] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5735–5744.
[68] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9317–9326.
[69] C. Han, R. Zheng, C. Gao, and N. Sang, "Complementation-reinforced attention network for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3433–3445, Oct. 2020.
[70] H. Lingxiao, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8450–8459.
[71] J. Lv, Z. Li, K. Nai, Y. Chen, and J. Yuan, "Person re-identification with expanded neighborhoods distance re-ranking," *Image Vis. Comput.*, vol. 95, Mar. 2020, Art. no. 103875.
[72] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8514–8522.
[73] S. Zhou, F. Wang, Z. Huang, and J. Wang, "Discriminative feature learning with consistent attention regularization for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8040–8049.
[74] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang, "Salience-guided cascaded suppression network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3300–3310.
[75] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6449–6458.
[76] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3186–3195.
[77] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep kronecker-product matching for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6886–6895.
[78] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 371–381.
[79] J. Zhou, B. Su, and Y. Wu, "Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2909–2918.

XU WANG received the graduation degree from the Department of Computer Science, Xiamen University, China. His current research interests include computer vision and image processing.

YAN HUANG is currently pursuing the graduation degree with the Department of Computer Science, Xiamen University, China. Her current research interests include machine learning and computer vision.

QICONG WANG received the Ph.D. degree in information and communication engineering from Zhejiang University, Hangzhou, China. He is currently an Associate Professor with the Department of Computer Science, Xiamen University, Xiamen, China. His research interests include computer vision, machine learning, and big data analytic.

YAN CHEN received the M.S. degree from Brunel University London, U.K. She is currently an Assistant Teacher with the College of Business and Management, Xiamen Huaxia University, Xiamen, China. Her current research interests include data analysis and modeling.

YEHU SHEN received the Ph.D. degree from Zhejiang University. He is currently an Associate Professor with the College of Mechanical Engineering, Suzhou University of Science and Technology, Suzhou, China. His research interests include computer vision, vSLAM, and machine learning. He is a Senior Member of the China Computer Federation (CCF).