Attention Deep Model with Multi-Scale Deep Supervision for Person Re-Identification

Di Wu, Chao Wang, Yong Wu and De-Shuang Huang, Senior Member, IEEE

Abstract—Person re-identification (PReID) has become a hot topic in computer vision due to it is an important part in surveillance systems. Many recent state-of-the-art PReID models are attention or multi-scale feature learning deep architectures. However, introducing attention mechanism may lead to some important feature information losing. Besides, most of the multi-scale models embed the multi-scale feature learning block into the feature extraction deep network, which reduces the efficiency of inference network. To address these issues, in this study, we introduce an attention deep architecture with multi-scale deep supervision for PReID. Technically, we contribute an inversion attention block to complement the attention block, and a novel multi-scale layer with deep supervision operator for training the backbone network. The proposed block and operator are only used for training, and discarded in test phase. Experimental have been performed on Market-1501, DukeMTMC-reID and CUHK03 datasets. All the experiment results demonstrate that our proposed model significantly outperforms other competitive state-of-the-art methods.

Index Terms—Person re-identification, deep learning, attention, multi-scale learning, deep supervision.

I. INTRODUCTION

PERS ON re-identification (PReID), which aims at re-identifying a specific pedestrian of interest taken by different cameras, or a single camera across different time in a camera network, has been extensively studied in recent years. Due to its significance in the application of intelligent video surveillance and security system, and the development of deep learning systems [1-3], and the established large-scale PReID datasets, the task has attracted a lot of attentions in computer vision community. However, the task is still difficult thanks to the large variations on captured pedestrian such as clothes, pose, illumination and uncontrolled complex background. To improve the PReID performance, extensive research has been reported in past years. These works can be divided into two categories: a) exploiting discriminative features to represent the person’s appearance, and b) learn suitable distance metrics for better computing the similarities between the paired person images. In the first category, the classical discriminative descriptors include the LOMO [4], ELF [5] and LBP [6]. In the second category, supervised learning with labeled image are used to obtain distance metric function, such as LMNN [7], ADMM [8], LFDA [9] and PCCA [10]. However, these methods extract features and learn distance metric separately, without consider their capabilities in a uniform structure.

Recently, with the successful development of the deep learning technology, many deep learning approaches take advantage of Convolutional Neural Network (CNN) to learn discriminative and robust deep features for PReID. These approaches usually are comprised of two components, i.e., a deep network and the objective functions. Early deep networks include VGGNet [11], ResNet [12] and DensNet [13]. Most recently, several attention-based deep models have been proposed, such as SENet [14], CBAM [15], BAM [16] and SKNet [17]. These models introduce attention module into the state-of-the-art deep architectures to learn the spatial information and relationship between channels. In general, the softmax scores produced by the attention module are multiplied by the original features to output the final emphasized features. As a part of whole features, the un-emphasized features are also important to enhance the discriminative ability of descriptors, especially when they contain the body information. We argue that the un-emphasized features should be treated as emphasized features to help to learn the final descriptors. However, few of current PReID works have considered this issue.

The idea of utilizing middle-level features of a deep framework has been investigated, and has proved to be useful for object detection [18] and segmentation [19]. The strategy is first used by [20] for PReID. They proposed a deep model that combines the embeddings of multiple convolutional network layers and trains them with deep supervision. Experimental results show the effectiveness of the strategy. Yet, they fused embeddings at the lower and higher layers for training and test, which reduces the efficiency of framework.

Recent works [20-22] have shown that multi-scale features learning can help to enhance the robust property of feature representations. Chen et al. [21] proposed a deep pyramidal feature learning framework which contains $m$ scale-specific

This work was supported by grants from the National Science Foundation of China, Nos. 61520106006, 61732012, 61861146002, 61772370, 61702371, 61672203, 61772357, and 61672382, China Postdoctoral Science Foundation Grant, Nos. 2017M611619, and by the Scientific & Technological Base and Talent Special Program of the Guangxi Zhuang Autonomous Region, GuiKe AD18126015 and by “BAGUI Scholar” Program of Guangxi Province of China. (Corresponding authors: De-Shuang Huang and Chao Wang)
branches for multi-scale deep feature learning. More specifically, each branch learns one scale in the pyramid. Besides, they used a scale-fusion branch to learn the complementary of combined multi-scale features. Qian et al. [22] explored different resolution levels of filters to learn pedestrian features at multiple locations and spatial scales. These methods have showed great benefits on performance in PReID. However, using multi-branches to obtain multi-scale information may increase the complexity of framework.

By reviewing the current PReID works, we find that the performance of deep model could be improved by introducing the following strategies: (1) attention mechanisms; (2) middle-level features for deep supervision; (3) multi-scale features learning. Nevertheless, using attention mechanisms can lead to the loss of important information. In addition, introducing the middle-level features to the final descriptors for deep supervision and adding the multi-scale features learning lower the efficiency of the model. In PReID domain, these issues are barely considered. Therefore, in this work, we propose an auxiliary training based deep multi-scale supervision attention model to deal with the above issues.

As shown in the Figure 1, the backbone of the proposed model is ResNet-50, which is used to extract different hierarchical deep features from the input image. The whole model is trained with a ranked triplet loss and four classification losses. To tackle the information loss caused by attention module, we introduce an inversion attention block into the middle-level features learning. The block outputs probability scores that are complementary to the softmax scores produced by attention block. We do the dot product between the probability scores and the original features, thus, the un-emphasized features become the emphasized features. Then these features are passed through the pooling and concatenated to perform the classification task together. To deal with the efficiency problem, we propose a multi-scale layer to learn multi-scale information before the deep supervision operation. The multi-scale layer consists of several lightweight single scale convolution kernels, which learns the multi-scale information from horizontal and vertical directions, respectively. The deep supervision operators with multi-scale information learning only assist the deep framework to learn in the training phase, which are discarded these operators in the test phase. In this way, the efficiency of the deep framework is improved.

The contributions of this work are summarized as followings:

1) We propose an inversion attention block to remedy the feature information loss caused by the attention block.

2) We propose a lightweight multi-scale features learning block to perform deep supervision, which helps to learn more discriminative deep feature with multi-scale information.

3) The proposed operators above are only performed in the training phase and discarded for testing, enhancing the effectiveness of inference network.

4) The proposed architecture outperforms other most recent state-of-the-art models on the three popular PReID datasets, i.e. Market-1501, DukeMTMC-reID and CUHK03.

II. RELATED WORKS

In this section, we review some related works in deep PReID domain. Then we introduce some recent attention deep models related to our method. Finally, some multi-scale deep models for PReID are briefly described.

A large number of deep learning based methods have been proposed for PReID. There mainly exist three types of deep PReID models, i.e. identification model [23-27], verification model [28-30] and distance metric learning model [31][32][33, 34]. Identification model formulates the PReID task as a classification problem. A fusion feature network was proposed by Wu et al. [23]. Through the supervision of identification loss, the model uses hand-crafted features to constrain the deep descriptors in the backpropagation phase. Verification model takes a pair of images as input and outputs a score to represent the similarity of the paired images. Li et al. [28] proposed a verification model named filter pairing neural network (FPNN) for PReID. Distance metric learning model makes the relative distance between positive pairs smaller than that of negative pairs. Alexander et al. [33] proposed a hard mining within a mini-batch for triplet loss. Wang et al. [35] proposed the Ranked List Loss (RLL) for building a set-based similarity structure by exploiting all instances in the gallery of deep metric learning. In this work, we combine the identification and Ranked List Loss to supervise the carefully designed deep architecture.

Attention is an important part in one person’s perception. Inspired by this, there have been several works try to incorporate attention mechanisms to the state-of-the-art deep classification models to improve the performance of them. Hu et al. [36] introduced a compact ‘Squeeze-and-Excitation’ module to exploit the relationships between inter-channels. In the compact module, the channel-wise attention was computed by the global average-pooling features. By introducing the spatial attention, Woo et al. [15] further proposed a convolution block attention module to simultaneously exploit the channel-wise and spatial attentions. They first used max-pooling and average-pooling operations to aggregate the channel information. Then these pooling features were concatenated and convolved to generate 2-dimensional spatial attention map. The attention models above have been proven to be efficient for improving the performance of deep network. However, introducing the attention mechanism may lead to feature information loss problem. Liu et al. [37] proposed a dual reverse attention networks for PReID. However, they perform the reverse attention operations on spatial and channel attentions separately. In the test phase, they used the concatenated reverse feature vectors to obtain the final descriptor. Different from their reverse attention modules, we perform the inversion attention block for spatial and channel attentions uniformly. Besides, our inversion attention block is only used for promoting the network to pay attention to the un-emphasized features, and is not used for testing, thus enhancing the timeliness of inference network.

Some works proposed to use multi-scale leaning for PReID. Liu et al. [38] proposed a multi-scale triplet deep architecture that learns deep features of a pedestrian at different scales.
Specifically, the architecture integrated shadow and deep networks to produce low-level and high-level appearance features from images, respectively. Wang et al. [20] introduced a deeply supervised method for PReID. They use the pooling feature map of each stage to produce an embedding for each stage. Then these embeddings were fused by using a weighted sum. Both these methods show potential for PReID. Yet, since the multi-scale modules are inserted to deep architecture for training and inference, the whole network becomes computationally. Aiming at this problem, we propose a lightweight multi-scale module with deep supervision operation to learn multi-scale information for the network.

III. PROPOSED METHOD

In this section, we present the carefully designed deep construct by first presenting the training network, and then describing the loss functions. Finally, the inference network is described.

A. Training Network

The overview of the proposed deep architecture is shown as Fig 1. We use ResNet-50 as the backbone network, in which the last spatial down-sampling operation, the original global average pooling and fully connected layers are removed. We append the average pooling layer and classification linear layer at the end of backbone network. As shown in Fig 1, we utilize the feature maps produced by stage1, stage2, stage3 and stage4 of the ResNet-50 to generate attention and inversion attention masks. Besides, for reducing the GPU memory cost of training network, we select the feature maps from stage2 and stage3 to perform multi-scale deep supervision operations. The detail structure of multi-scale layer is shown in Figure 2. The whole architecture is supervised by five losses functions: four classification losses and one triplet loss.

B. Attention Block

Enlighten by the works of [15] and [16], the attention block adopted in our architecture is consisted of channel attention and spatial attention. The channel attention outputs a set of weights for different channels while the spatial attention concentrates on the informative part.

Channel Attention: The channel attention block contains an average pooling layer and two linear layers. The feature maps \( \mathbf{M} \) produced by the end of each stage are first passed through the average pooling operator as below:

\[
\mathbf{M}_c = \text{AvgPool}(\mathbf{M})
\]

where \( \mathbf{M} \in \mathbb{R}^{C \times W \times H} \), \( \mathbf{M}_c \in \mathbb{R}^{C \times 1 \times 1} \).

Then the two linear layers with batch normalization are used for estimating attention across channels from \( \mathbf{M}_c \). To reduce the overhead of parameters, the output size of the first linear layer is set to \( C / r \), in which \( r \) is the reduction ratio. To
restore the channel number, the output of the second linear layer is set to \( C \). After the linear layers, a batch normalization layer is appended to adjust the scale of the spatial attention output. Thus, the channel attention \( ATT_C \) can be computed as:
\[
ATT_C = BN(linear1(linear2(M_c)))
\]  
where \( linear1, linear2 \) and \( BN \) represent the first linear layer, the second layer and batch normalization layer, respectively.

**Spatial Attention:** Spatial attention block is used to emphasize features at different spatial positions. The block contains two reduction layers and two convolutions layers. Through the first reduction layer, the dimension of feature maps \( M \in \mathbb{R}^{C \times W \times H} \) is reduced to \( M_S \in \mathbb{R}^{C' \times W' \times H'} \) with \( 1 \times 1 \) kernel size. Then the \( M_S \) is passed through the two convolution layers with \( 3 \times 3 \) kernel size. Finally, the features are further reduced to \( \mathbb{R}^{1 \times W' \times H'} \) by using the second reduction layer with \( 1 \times 1 \) kernel size. Similar to the channel attention block, we apply the batch normalization operation at the end of the spatial attention block. The spatial attention \( ATT_S \) can be written as:
\[
ATT_S = BN(Reduction2(Conv2(Conv1(M_c))))
\]  
where \( Conv1 \) and \( Conv2 \) represent the two convolution layers, \( Reduction2 \) is the second reduction layer.

**Combine the two attentions:** Channel attention and Spatial attention are combined by:
\[
ATT = \sigma(ATT_C \times ATT_S)
\]  
where \( ATT \) is the output of the whole attention block, and \( \sigma \) is the Sigmoid function.

**C. Inversion Attention**

The attention block generates one probability mask to jointly emphasize the channels, and features in different spatial locations. However, through this operation, some channels and features are suppressed. We argue that these suppressed channels and features may be informative, which helps to enhance the discriminative ability of the final descriptors. Based on this assumption, we introduce an inversion attention to complement the attention block. The output of the inversion attention in our model can be expressed as:
\[
ATT_R = 1 - \sigma(ATT_C \times ATT_S)
\]

We use the feature maps generated by each stage to do dot product with \( ATT_R \). In this way, these suppressed channels and features become emphatic. Then these emphasized feature maps at each stage are passed through a pooling layer and concatenated to perform classification task.

**D. Deep Supervision with Multi-Scale Learning**

From Fig 1, we use the feature maps at the stage 2 and stage 3 to perform deep supervision operations, respectively. The deep supervision is helpful for learning the discriminative descriptors. Note that we add the deep supervision operations behind the attention block, which is conducive to get more accurate attention maps. Besides, we introduce a multi-layer to acquire the multi-scale information for deep supervision. The structure of multi-layer is shown as Figure 2, we first partition the channels into equal four groups. Then the four groups are passed through four convolution operations with kernel sizes of \( 1 \times 3, 3 \times 1, 1 \times 5 \) and \( 5 \times 1 \), respectively. After the convolution operations, the four groups are concatenated into one group. The reasons of why we choose one dimensional convolution as are as follows:

- a) One dimensional convolution has less parameter, and reduces GPU memory consumption.
- b) One dimensional convolution operations can learn the pedestrian features from horizontal and vertical directions, respectively, which adapts the human visual perception.

**E. Loss Functions**

There have five loss functions in our model, *i.e.* four ID losses and one triplet loss. We employ the ranked list loss (RLL) proposed by [40] for metric learning and smoothing cross entropy loss proposed by [41] for classification. The overall loss is the sum of them.

**Ranked List Loss:** We use the RLL to supervise the branch-2. The goal of RLL is to make the distance between negative samples larger than a threshold \( \alpha \), while the distance between positive samples closer than a threshold \( \alpha - m \), where \( m \) is a margin. The function can be written as:
\[
L_m(x_i, x_j, f) = (1 - y_{ij})[(\alpha - d_{ij})_+ + y_{ij}[d_{ij} - (\alpha - m)]_+]
\]  
where \( y_{ij} = 1 \) denotes \( x_i \) and \( x_j \) are the same person. \( y_{ij} = 0 \), otherwise. \( d_{ij} \) represents the Euclidean distance function, which is formulated as \( d_{ij} = \|f(x_i) - f(x_j)\| \), \( f(x_i) \) is the embedding function.

The nontrivial positive sample set can be displayed as:
\[
P_{\alpha}^+ = \{x_i^j | j \neq i, d_{ij} > \alpha - m\}
\]
The nontrivial negative sample set can be represented as:

\[ N_{c, l}^\prime = \{ x_i^k | k \neq c, d_{ij} < \alpha \} \]  

(8)

To pull all the nontrivial positive samples closer, we should minimize the following objective function:

\[ L_P(x_i^c; f) = \frac{1}{|P_{c, l}|} \sum_{x_i^c \in P_{c, l}} L_m(X_i^c, X_i^c; f) \]  

(9)

In order to push the distance between nontrivial negative samples larger than the threshold \( \alpha \), the following objective function should be minimized:

\[ L_N(x_i^c; f) = \sum_{x_i^c \in [N_{c, l}]} \frac{w_{ij}}{\sum_{x_i^c \in [N_{c, l}]} w_{ij}} L_m(X_i^c, X_i^c; f) \]  

(10)

where \( w_{ij} \) is the weight of nontrivial negative samples.

Then the function of RLL can be written as:

\[ L_{RLL}(x_i^c; f) = L_P(x_i^c; f) + \lambda L_N(x_i^c; f) \]  

(11)

in which the \( \lambda \) is the balance parameter, and we set it to 1.

**Smoothing Cross-Entropy Loss:** In order to alleviate the effect of the over-fitting problem, we use label smoothing with cross-entropy loss to supervise the branch-1, branch-3, branch-4 and branch-5. Label smoothing function can be defined as:

\[ q_i = \begin{cases} 1 - \frac{(N-1)\varepsilon}{N} & \text{if } i = y \\ \varepsilon \frac{n}{N} & \text{otherwise} \end{cases} \]  

(12)

where \( y \) is the true label, \( i \) represents the predicted label, \( N \) is the number of training samples, \( \varepsilon \) is a constant and we set it to 0.1. Then the function of smoothing cross-entropy loss is formulated as:

\[ L_{1D} = \sum_{i=1}^{N} -q_i \log (p_i) \]  

(13)

where \( p_i \) denotes as the prediction logits of class \( i \).

The overall loss function of the architecture can be written as:

\[ L = \lambda_1 L_{RLL} + \lambda_2 L_{1D1} + \lambda_3 L_{1D2} + \lambda_4 L_{1D3} + \lambda_5 L_{1D4} \]  

(14)

where \( L \) is the total loss, \( \lambda_1, \lambda_2, \lambda_3, \lambda_4 \) and \( \lambda_5 \) are the balance coefficients.

**F. Inference Network**

Our inference network is quite efficient and simple. As shown in Fig 3, in the test phase, we discard the multi-scale deep supervision, the inversion attention and the triplet branches, i.e. branch-1, branch-2, branch-4 and branch-5, and only use branch-3 to predict.

IV. **EXPERIMENTS**

A. **Datasets**

We perform the experiments on three popular PReID datasets, i.e. Market-1501 [42], CUHK03 [43] and DukeMTMC-reID [44]. The brief introductions of the datasets are presented as below:

**Market-1501 dataset:** It contains 32643 images with 1501 pedestrians captured by by at least two cameras and at most six cameras from a supermarket. The training set and testing set contain 12936 images of 751 IDs and 19732 images of 750 IDs, respectively.

**CUHK03 dataset:** The dataset contains 14097 images with 1467 pedestrians. It provides two bounding boxes detection settings. One manually annotated by human and the other automatically annotated by a detector. We conduct the experiments both on the two settings. Similar to [45], we divide the dataset into training set with 767 pedestrians and test set with 700 pedestrians.

**DukeMTMC-reID dataset:** It consists of 36411 annotated boxes with 1812 pedestrians captured by eight cameras. Among the 1812 pedestrians, 1404 pedestrians appeared in more than two camera views, and the rest of pedestrians are treated as distractor identifications. The training set of this dataset consists of 16522 images of 702 pedestrians and the test set contains 17661 gallery images and 2228 query images.

B. **Evaluation Metrics**

Cumulative match characteristic (CMC) and mean average precision (mAP) are used as evaluation metrics to estimate the performance of our model. Besides, we report the Rank-1 and Rank-5 results. Both experiments of the datasets are conducted under single query mode. Note that we don’t use re-ranking in this work.

C. **Implementation Details**

The proposed architecture is implemented on Pytorch. We use the ResNet-50 with pre-trained parameters on ImageNet as the backbone network. The hyper parameter reduction ratio \( r \)

---

*Fig. 3. The inference network.*
of attention block is set to 16. The commonly used data augmentation methods in PReID are followed. We resize all the image into 384×128. Random erasing [46] and horizontal random flipping are used for training. The batch size contains 64 images with 16 identities, in which each identity has 4 images. We train the architecture for 120 epochs. The $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ and $\lambda_5$ described in Eq. (14) are set to 0.4, 0.1, 0.03 and 0.03, respectively. We utilize adaptive moment estimation with an initial learning rate ($lr$) of $3.5 \times 10^{-5}$ to optimize the five loss functions. Following the work of [39], the $lr$ is then updated as the following rules:

$$lr(t) = \begin{cases} 
3.5 \times 10^{-5} \times \frac{1}{10} & \text{if } t \leq 10 \\
3.5 \times 10^{-4} & \text{if } 10 < t \leq 40 \\
3.5 \times 10^{-5} & \text{if } 40 < t \leq 70 \\
3.5 \times 10^{-6} & \text{if } 70 < t \leq 120 
\end{cases}$$

(15)

All the experiments run on two TITAN XP GPUs.

D. Comparison with Related Methods

The proposed method is compared with the following state-of-the-art models, which includes PNGAN [47], PABR [48], PCB+RPP [49], SGGN [50], MGN [51], G2G [52], SPReID [53], IANet [54], CASN [55], OSNet [56], BDB+Cut[57], P2-Net [58], and so on.

**Evaluation on market-1501 dataset.** The comparison results between our model and the state-of-the-arts on market-1501 dataset are shown in Table I. From Table I, we can find that the proposed method outperforms the other competing models. Compared to Mancs which also utilizes attention and deep supervision operations, our model increases the mAP by +6.7% and Rank-1 by +2.4%, respectively. The proposed model achieves mAP= 89.0%, Rank-1 accuracy= 95.5% and Rank-5 accuracy= 98.3% under single query mode, which validates the efficiency of it.

| Method      | Publication | mAP | R-1 | R-5 |
|-------------|-------------|-----|-----|-----|
| PNGAN [47]  | ECCV 18     | 72.6| 89.4| --  |
| PABR [48]   | ECCV 18     | 76.0| 90.2| 96.1|
| PCB+RPP [49]| ECCV 18     | 81.6| 93.8| 97.5|
| SGGN [50]   | ECCV 18     | 82.8| 92.3| 96.1|
| Mancs [59]  | ECCV 18     | 82.3| 93.1| --  |
| MGN [51]    | MM18        | 86.9| 95.7| --  |
| FDGAN [60]  | NeurIPS 18  | 77.7| 90.5| --  |
| DaRe [20]   | CVPR 18     | 76.0| 89.0| --  |
| PSE [61]    | CVPR 18     | 69.0| 87.7| 94.5|
| G2G [52]    | CVPR 18     | 82.5| 92.7| 96.9|
| DeepCRF [62]| CVPR 18     | 81.6| 93.5| 97.7|
| SPReID [53] | CVPR 18     | 81.3| 92.5| 97.2|
| KPM [63]    | CVPR 18     | 75.3| 90.1| 96.7|
| AANet [64]  | CVPR 19     | 83.4| 93.9| --  |
| CAMA [65]   | CVPR 19     | 84.5| 94.7| 98.1|
| IANet [54]  | CVPR 19     | 83.1| 94.4| --  |
| DGNet [66]  | CVPR 19     | 86.0| 94.8| --  |
| CASN [55]   | CVPR 19     | 82.8| 94.4| --  |

**Table I**

**Comparison results on market-1501.**

| Method      | Publication | R-1 | R-5 |
|-------------|-------------|-----|-----|
| Ours        | --          | 78.8| 75.3|

**Evaluation on CUHK03 dataset.** For this dataset, we adopt the new protocol introduced by [45], in which 767 pedestrians are utilized for training and the rest of 700 pedestrians are used for test. The comparison results on CUHK03_detected and CUHK03_labeled settings are shown in Table II and Table III, respectively. We report mAP and Rank-1 accuracy under single query mode. From the two Tables, we can find that the proposed method also beats all other compared state-of-the-art approaches, showing the efficiency of our method. Compared with Mancs, our method improves the mAP and Rank-1 accuracy by at least 13 percent, which further demonstrates the advantages of the proposed model.

| Method      | Publication | R-1 | mAP |
|-------------|-------------|-----|-----|
| MGN [51]    | MM18        | 66.8| 66.0|
| PCB+RPP [49]| ECCV 18     | 63.7| 57.5|
| Mancs [59]  | ECCV 18     | 65.5| 60.5|
| DaRe [20]   | CVPR 18     | 63.3| 59.0|
| CAMA [65]   | CVPR 19     | 66.6| 64.2|
| CASN [55]   | CVPR 19     | 71.5| 64.4|
| OSNet [56]  | ICCV 19     | 72.3| 67.8|
| Auto-ReID [68]| ICCV 19    | 73.3| 69.3|
| BDB+Cut [57]| ICCV 19     | 76.4| 73.5|
| MHN-6 [69]  | ICCV 19     | 71.7| 65.4|
| P2-Net [58] | ICCV 19     | 74.9| 68.9|

**Table II**

**Comparison results on CUHK03_detected.**

| Method      | Publication | R-1 | mAP |
|-------------|-------------|-----|-----|
| Ours        | --          | 81.0| 78.2|
Evaluation on the DukeMTMC-reID dataset. As shown in Table IV. Our proposed method achieves 79.2%/89.4% in mAP/Rank-1 on the DukeMTMC-reID dataset. Compared with the recent methods, we achieve the best results in mAP and Rank-1, which exceeds the state-of-the-art method MHN-6 by +2% and +0.3%, respectively.

### TABLE IV
**COMPARISON RESULTS ON DUKEMTMC-REID.**

| Method      | Publication | mAP   | R-1     | R-5     | R-10    |
|-------------|-------------|-------|---------|---------|---------|
| G2G [52]    | CVPR 18     | 66.4  | 80.7    | 88.5    | 90.8    |
| DeepCRF [62] | CVPR 18     | 69.5  | 84.9    | 92.3    | --      |
| SPReID [53] | CVPR 18     | 71.0  | 84.4    | 91.9    | 93.7    |
| PABR [48]   | ECCV 18     | 64.2  | 82.1    | 90.2    | 92.7    |
| PCB+RPP [49]| ECCV 18     | 69.2  | 83.3    | 90.5    | 95.0    |
| SGGNN [50]  | ECCV 18     | 68.2  | 81.1    | 88.4    | 91.2    |
| Mancs [59]  | ECCV 18     | 71.8  | 84.9    | --      | --      |
| MGN [51]    | MM18        | 78.4  | 88.7    | --      | --      |
| AANet [64]  | CVPR 19     | 74.3  | 87.7    | --      | --      |
| CAMA [65]   | CVPR 19     | 72.9  | 85.8    | --      | --      |
| IANet [54]  | CVPR 19     | 73.4  | 87.1    | --      | --      |
| DGNet [66]  | CVPR 19     | 74.8  | 86.6    | --      | --      |
| CASN [55]   | CVPR 19     | 73.7  | 87.7    | --      | --      |
| OSNet [56]  | ICCV 19     | 74.8  | 86.6    | --      | --      |
| Auto-ReID [68]| ICCV 19    | 75.1  | 88.5    | --      | --      |
| BBD+Cut [57]| ICCV 19     | 76.0  | 89.0    | --      | --      |
| P²-Net [58] | ICCV 19     | 73.1  | 86.5    | 93.1    | 95.0    |
| MHN-6 [69]  | ICCV 19     | 77.2  | 89.1    | 94.6    | 96.5    |
| **Ours**    | --          | **79.2** | **89.4** | **94.7** | **96.0** |

### E. Ablation Studies and Discussions

We further implement some extra experiments to evaluate the effectiveness of each block of the our model. All the ablation experiments are performed on CUHK03 labeled dataset under single query mode. The details of experiment results of ablation studies are shown as below:

**Effectiveness of the inversion attention block.** In this setting, we discard the inversion attention and named the discarded model as Our\textsubscript{inversion}. The experiment results of the model on CUHK03 labeled dataset are shown in Table V. From Table V, we can observe that when discard the inversion attention block, the performance of the model is decreased. More specifically, without inversion attention, the mAP and Rank-1 accuracy are reduced by -1.5% and -3.7%, respectively.

To further prove the effectiveness of the inversion attention block, some feature maps generated by Our\textsubscript{inversion} and Our models are presented in Fig 4. From Fig 4, we find that the fine details of feature maps are enriched when introducing the inversion attention block.

**Effectiveness of the multi-scale deep supervision block.** To verify the effectiveness of the multi-scale deep supervision block, we use the model that discards branch-4 and branch-5 to implement the experiment. As shown in Table VI, when introducing the multi-scale deep supervision block, the performance of the discarded model Our\textsubscript{supervision} is improved, increasing the mAP and Rank-1 accuracy by +1.3% and 1.9%, respectively.

### TABLE V
**THE EFFECTIVENESS OF ADVERSE ATTENTION BLOCK**

| Model       | mAP  | R-1   | R-5   |
|-------------|------|-------|-------|
| Our\textsubscript{inversion} | 76.7 | 77.3 | 91.3 |
| Our          | **78.2** | **81.0** | **92.0** |

### TABLE VI
**THE EFFECTIVENESS OF MULTI-SCALE DEEP SUPERVISION BLOCK**

| Model       | mAP  | R-1   | R-5   |
|-------------|------|-------|-------|
| Our\textsubscript{supervision} | 76.9 | 79.1 | 91.6 |
| Our          | **78.2** | **81.0** | **92.0** |

### V. Conclusion

In this study, we introduce an attention deep architecture with multi-scale deep supervision for person re-identification. We first design the inversion attention block to complement the attention block, then we introduce a multi-scale deep supervision block to learn features with multi-scale information as well as rectify the attention block. The experiment results on the three large datasets show that the proposed model outperforms the previous state-of-the-art methods. What’s more, in this work, we only divided the channels into four groups in the multi-scale layers, we believe...
if we divide the channels into finer groups, the accuracy of PReID can be further improved. In the future, we would pay more attention to the cross-modality issue for PReID.

REFERENCES

[1] D. S. Huang and J. X. Du, "A Constructive Hybrid Structure Optimization Methodology for Radial Basis Probabilistic Neural Networks," *IEEE Transactions on Neural Networks*, vol. 19, pp. 2099-115, 2008.

[2] D.-S. HUANG, "Radial basis probabilistic neural networks: model and application," *International Journal of Pattern Recognition & Artificial Intelligence*, vol. 13, pp. 1083-1101, 1999.

[3] D.-S. Huang, "Systematic theory of neural networks for pattern recognition," *Publishing House of Electronic Industry of China, Beijing*, vol. 201, 1996.

[4] R. Zhao, W. Ouyang, and X. Wang, "Learning Mid-level Filters for Person Re-identification," in *computer vision and pattern recognition*, 2014, pp. 144-151.

[5] D. Gray and H. Tao, "Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features," in *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I*, 2008.

[6] F. Xiong, M. Gou, O. Camps, and M. Szaier, "Person Re-Identification Using Kernel-Based Metric Learning Methods," 2014.

[7] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian Recognition with a Learned Metric," in *Asian Conference on Computer Vision*, 2010, pp. 501-512.

[8] N. Martinel, A. Das, C. Micheloni, and A. K. Roychowdhury, "Temporal Model Adaptation for Person Re-Identification," *european conference on computer vision*, pp. 858-877, 2016.

[9] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghosian, "Local Fisher Discriminant Analysis for Pedestrian Re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3318-3325.

[10] F. Jurie and A. Mignon, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2666-2672.

[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Computer Science*, 2014.

[12] K. He, X. Zhang, S. Ren, and S. Jian, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2016.

[13] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *CVPR*, 2017.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," *arXiv: Computer Vision and Pattern Recognition*, 2017.

[15] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *european conference on computer vision*, 2018, pp. 3-19.

[16] J. Park, S. Woo, J. Lee, and I. S. Kweon, "RAN: Bottleneck Attention Module," *arXiv: Computer Vision and Pattern Recognition*, 2018.

[17] X. Li, W. Wang, X. Hu, and J. Yang, "Selective Kernel Networks," in *computer vision and pattern recognition*, 2019, pp. 510-519.

[18] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, pp. 640-651, 2014.

[20] Y. Wang, L. Wang, Y. You, X. Zou, V. S. Chen, S. Li, et al., "Resource Aware Person Re-identification Across Multiple Resolutions," in *computer vision and pattern recognition*, 2018, pp. 8042-8051.

[21] Y. Chen, X. Zhu, S. Gong, and IEEE, "Person Re-identification by Deep Learning Multi-Scale Representations," in *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2017.

[22] X. Qian, Y. Fu, T. Xiang, Y. Jiang, and X. Xue, "Leader-based Multi-Scale Attention Deep Architecture for Person Re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-1, 2019.

[23] S. Wu, Y. C. Chen, X. Li, A. C. Wu, J. J. You, and W. S. Zheng, "An enhanced deep feature representation for person re-identification," in *Applications of Computer Vision*, 2016, pp. 1-8.

[24] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification," in *Computer Vision and Pattern Recognition*, 2016, pp. 1249-1258.

[25] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," *arXiv preprint arXiv:1703.07220*, 2017.

[26] D. Wu, S. Zheng, C. Yuan, and D. Huang, "A deep model with combined losses for person re-identification," *Cognitive Systems Research*, vol. 54, pp. 74-82, 2019.

[27] D. Wu, H. Yang, D. Huang, C. Yuan, X. Qin, Y. Zhao, et al., "Omnidirectional Feature Learning for Person Re-Identification," *IEEE Access*, vol. 7, pp. 28402-28411, 2019.

[28] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep Filter Pairing Neural Network for Person Re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152-159.

[29] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Computer Vision and Pattern Recognition*, 2015, pp. 3908-3916.

[30] L. Wu, C. Shen, and A. v. d. Hengel, "Personnet: Person re-identification with deep convolutional neural networks," *arXiv preprint arXiv:1601.07255*, 2016.

[31] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 81, pp. 2993-3003, 2018.

[32] 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 *Computer Vision and Pattern Recognition*, 2016, pp. 1335-1344.

[33] A. Hermans, L. Beyer, and B. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," *arXiv preprint arXiv:1703.07737*, 2017.

[34] D. Wu, S. Zheng, W. Bao, X. Zhang, C. Yuan, and D. Huang, "A novel deep model with multi-loss and efficient training for person re-identification," *Neurocomputing*, vol. 324, pp. 69-75, 2019.

[35] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. A. Robertson, "Ranked List Loss for Deep Metric Learning," in *computer vision and pattern recognition*, 2019, pp. 5207-5216.

[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," *arXiv: Computer Vision and Pattern Recognition*, 2017.

[37] S. Liu, L. Qi, Y. Zhang, and W. Shi, "Dual reverse attention networks for person re-identification," in *international conference on image processing*, 2019.

[38] J. Liu, Z. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, et al., "Multi-Scale Triplet CNN for Person Re-Identification," in *acm multimedia*, 2016, pp. 192-198.

[39] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of Tricks and a Strong Baseline for Deep Person Re-Identification," in *computer vision and pattern recognition*, 2019, pp. 0-0.

[40] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. A. Robertson, "Ranked List Loss for Deep Metric Learning," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *computer vision and pattern recognition*, 2016, pp. 2818-2826.

[42] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable Person Re-identification: A Benchmark," in *international conference on computer vision*, 2015, pp. 1116-1124.

[43] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep Filter Pairing Neural Network for Person Re-identification," in *computer vision and pattern recognition*, 2014, pp. 152-159.

[44] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled Samples Generated by GN: Improve the Person Re-identification Baseline in Vitro," *international conference on computer vision*, pp. 3774-3782, 2017.

[45] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking Person Re-identification with k-Reciprocal Encoding," in *computer vision and pattern recognition*, 2017, pp. 3652-3661.
[46] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," *arXiv: Computer Vision and Pattern Recognition*, 2017.

[47] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, et al., "Pose-Normalized Image Generation for Person Re-identification," in *European conference on computer vision*, 2018, pp. 661-678.

[48] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-Aligned Bilinear Representations for Person Re-identification," in *European conference on computer vision*, 2018, pp. 418-437.

[49] 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 *European conference on computer vision*, 2018, pp. 501-518.

[50] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person Re-identification with Deep Similarity-Guided Graph Neural Network," *arXiv: Computer Vision and Pattern Recognition*, 2018.

[51] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning Discriminative Features with Multiple Granularities for Person Re-Identification," *ACM Multimedia*, pp. 274-282, 2018.

[52] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang, "Deep Group-Shuffling Random Walk for Person Re-identification," in *Computer vision and pattern recognition*, 2018, pp. 2265-2274.

[53] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kanasak, and M. Shah, "Human Semantic Parsing for Person Re-identification," in *Computer vision and pattern recognition*, 2018, pp. 1062-1071.

[54] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-Aggregation Network for Person Re-identification."

[55] M. Zheng, S. Karanam, Z. Wu, and R. J. Rudke, "Re-Identification With Consistent Attentive Siamese Networks," in *Computer vision and pattern recognition*, 2019, pp. 5735-5744.

[56] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omniscale Feature Learning for Person Re-Identification," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[57] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch DropBlock Network for Person Re-identification and Beyond," *arXiv: Computer Vision and Pattern Recognition*, 2018.

[58] J. Guo, Y. Yuan, L. Huang, C. Zhang, J. Yao, and K. Han, "Beyond Human Parts: Dual Part-Aligned Representations for Person Re-Identification," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[59] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A Multi-task Attentional Network with Curriculum Sampling for Person Re-identification," in *European conference on computer vision*, 2018, pp. 384-400.

[60] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, et al., "FD-GAN: Pose-guided Feature Distilling GAN for Robust Person Re-identification," in *Neural information processing systems*, 2018, pp. 1222-1233.

[61] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A Pose-Sensitive Embedding for Person Re-Identification with Expanded Cross Neighborhood Re-Ranking," *arXiv: Computer Vision and Pattern Recognition*, 2017.

[62] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group Consistent Similarity Learning via Deep CRF for Person Re-identification," in *Computer vision and pattern recognition*, 2018, pp. 8649-8658.

[63] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-End Deep Kronecker-Product Matching for Person Re-identification," in *Computer vision and pattern recognition*, 2018, pp. 6886-6895.

[64] C. Tay, S. Roy, and K. Yap, "A2ANet: Attribute Attention Network for Person Re-Identifications," in *Computer vision and pattern recognition*, 2019, pp. 7134-7143.

[65] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards Rich Feature Discovery With Class Activation Maps Augmentation for Person Re-Identification," in *Computer vision and pattern recognition*, 2019, pp. 1389-1398.

[66] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint Discriminative and Generative Learning for Person Re-identification," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[67] H. Cai, Z. Wang, and J. Cheng, "Multi-Scale Body-Part Mask Guided Attention for Person Re-Identification," in *Computer vision and pattern recognition*, 2019, pp. 0-0.