Learning Multi-Scale Features and Batch-Normalized Global Features for Person Re-Identification

ZONGJING CAO\textsuperscript{1} AND HYO JONG LEE\textsuperscript{1,2}, (Member, IEEE)

\textsuperscript{1}Department of Computer Science and Engineering, Jeonbuk National University, Jeonju 54896, South Korea
\textsuperscript{2}Center for Advanced Image and Information Technology, Jeonbuk National University, Jeonju 54896, South Korea

Corresponding author: Hyo Jong Lee (hlee@jbnu.ac.kr)

This was supported by the National Research Foundation of Korea through the Basic Science Research Program funded by the Ministry of Education under Grant GR 2019R1D1A3A03103736.

\textbf{ABSTRACT} In recent years, person re-identification based on deep learning approaches has made great progress and achieved good results. However, many of the latest network design methods, which usually deploy ResNet or SENet as the backbone network, were originally designed for classification tasks. Since the person re-identification task is essentially different from the classification task, the structure of the backbone network should be modified accordingly. In this paper, we propose a retrieval network based on a multi-scale backbone architecture, which is specifically suitable for the person re-identification task. By constructing hierarchical residual-like connections within a single residual block, the model learns multi-scale discriminative features of pedestrian images. Unlike many state-of-the-art methods that use complex network structures and concatenate multi-branch features, our proposed retrieval network is implemented using only global features, simple triplet loss, and softmax with cross-entropy loss. The results of extensive experiments show that the proposed network has stronger fine-grained pedestrian representation ability, leading to performance gains for person re-identification tasks. Our proposed network achieves a rank-1 accuracy of 96.03\% on the Market-1501 and 92.11\% on DukeMTMC-reID datasets while only using global features.

\textbf{INDEX TERMS} Person re-identification, convolutional neural networks, deep learning.

\section{I. INTRODUCTION}

Person re-identification (re-ID) is usually regarded as a sub-task of digital image retrieval. Given an image (query) of a pedestrian captured from one surveillance camera, person re-ID is the process of identifying the same person from images (gallery) taken from a different surveillance camera with non-overlapping fields of view [1]. Due to its application in an intelligent camera surveillance system, person re-ID is currently one of the most popular research topics in the pattern recognition and computer vision communities. Although the academic and industrial attention paid to person re-ID tasks has been increasing, it remains a challenging task due to similarities in pedestrian clothing, illumination variation, the low resolution of surveillance cameras, partial occlusion of pedestrians, and intra-class variation across cameras [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Hai Wang\textsuperscript{1}.

Recently, the convolutional neural networks (CNN) have shown great potential for learning feature embeddings and distance metric learning in computer vision tasks. Person re-ID based on CNN models for feature representation has made some progress and achieved good performance. However, many of the latest re-ID models are designed using ResNet [3] or SENet [4] as backbone networks. These models were originally designed for classification tasks. Xie \textit{et al.} [5] proved that the essentials of image retrieval and image classification are the same because both tasks can be solved by measuring the similarity between images. However, there are still a few differences between the classification task and the person re-ID task. First, there are differences in the input images. In the image classification task, the input images of the network are different objects, such as cars and ships, while in the person re-ID task, the inputs of the network are all pedestrian images. Second, the outputs of the models are different. The classification task aims to determine the generic
class or category of the query, while the goal of the retrieval task is to rank many candidates according to their relevance to the query. Third, the algorithms differ. The classification task aims to maximize the between-class distance and minimize the within-class distance, while person re-ID mainly focuses on improving the representation capability (representation learning) of features extracted from the CNN models and learning better similarity measures (metric learning). Broadly speaking, classification aims to group candidates into a highly general class, while the re-ID task has the much more challenging aim of segregating a single individual from all other individuals. On the other hand, for image classification tasks, the training images and test images have the same classes, and the testing images fall into these predefined classes. However, for person re-ID tasks, testing identities are not available because they do not have overlap with training identities and the training and testing images only consist of pedestrians [1].

Deep learning-based re-ID has received great attention in recent years and has become a research hotspot in the re-ID community. To improve the representation ability of the features extracted from CNN, researchers in this field have designed hundreds of different CNNs. However, many state-of-the-art models are designed to connect multi-branch features with complex network structures. For example, to achieve a higher accuracy, researchers combine multiple local features in the CNN or utilize the semantic information from pose estimation or segmentation models. Such complex network structures not only require substantial computational training overhead, but also greatly reduce the speed of the retrieval process [5].

In this paper, we propose a new retrieval network called Re-Net that is suitable for the person re-ID task. Unlike many state-of-the-art methods that use complex networks to represent multiscale pedestrian features in a layer-wise manner, we have improved the pedestrian representation ability of CNN models at a more granular level. The proposed Re-Net network is implemented using only global features, simple triplet loss, and identification loss. During the training process, the network not only considered the classification label information, but also considered the distance between the images in the class. Furthermore, we proposed a feature normalization network (FN-Net) to overcome the problem in which the targets of the two losses are inconsistent in the embedding space. Recently, batch normalization and dropout are commonly used methods to solve the overfitting problem of the CNN models. Dropout randomly discards the output of each hidden neuron with a probability during the training process. In FN-Net, we added a dropout layer to overcome the overfitting problem of the CNN model. The results of extensive experiments on three popular person re-ID benchmarks datasets show that the Re-Net network has a stronger fine-grained pedestrian representation ability, leading to performance gains on person re-ID tasks.

In comparison with existing works, the main contributions of this work are summarized as follows:

1) We propose a new Re-Net network based on a multi-scale backbone architecture for person re-ID task. More specifically, we design a feature extraction network (FE-Net) to obtain multi-scale representations of pedestrian features. By increasing the receiving range of each residual module, the FE-Net improves the representation ability of pedestrian multi-scale features at a more fine-grained level, which is specifically suitable for the person re-identification task.

2) We design an FN-Net to separate the identification loss and triplet loss into two different feature spaces. By using two different features to calculate identification loss and triplet loss, the FN-Net effectively solves the problem of inconsistency between the targets of the two losses in the embedding space.

3) We evaluate the performance of the Re-Net network on three large-scale benchmark datasets, which achieves 92.11% rank-1 and 90.10% mean average precision on the DukeMTMC-reID datasets while only using global features.

The rest of this paper is organized as follows: We first introduce the related works of person re-ID in Section II. Then, in Section III, we describe the proposed re-ID network structure in detail. In Section IV, we introduce the experiments, including the three large-scale benchmark datasets, evaluation protocol, and implementation details. Section V provides a comparison of the experiments on three benchmark datasets and their results. The conclusion and future works are provided in Section VI.

II. RELATED WORKS

In this section, we first review handcrafted systems for person re-ID and then introduce person re-ID algorithms based on deep learning methods.

A. TRADITIONAL METHODS FOR PERSON RE-ID

In general, the algorithm of person re-ID tasks consists of two components: (1) the method of extracting features from the input pedestrian images; (2) the metric to compare the similarity of these features in the feature space. Therefore, the typical person re-ID methods can be roughly divided into feature-based methods and metric-based methods [2]. The goals of feature-based methods are to find an effective descriptor for pedestrian representations, while the metric-based methods focus on learning an effective feature metric to reduce the distance of images corresponding to the same pedestrian and increase the distance between different pedestrians. Research on person re-ID mainly concentrates on how to either find an improved set of features [7] or find an improved similarity metric for comparing features [9].

Before deep learning methods dominated the re-ID research community, many hand-crafted algorithms were developed to learn the global and local features of pedestrians. Traditional person re-ID methods mainly concentrate on how to either develop a new feature representation or learn a new
distance metric [11]. For re-ID, it is necessary to address the changes in lighting, pose, and viewpoint of pedestrian images across cameras. For these purposes, color histograms [12], local binary patterns [10], Gabor features [9], color names [7], scale-invariant feature transforms [13] are commonly used features for person re-ID.

The main idea of the metric learning-based method is to find the mapping from the feature space to the new space so that the feature vectors from the same pedestrian image are closer than feature vectors from the different pedestrian images. Metric learning techniques that have been applied to person re-ID include Mahalanobis metric learning [14], locally adaptive decision functions [15], marginal Fisher analysis [12], and local Fisher discriminant analysis [12].

B. PERSON RE-ID BASED ON DEEP LEARNING METHODS

After the success of AlexNet [14] in ILSVRC 2012, more and more methods based on deep learning have been proposed for person re-ID. Different from the traditional methods, feature extraction is implicitly learned by CNN models instead of through handcrafted systems.

The CNN models commonly employed in the community can be broadly divided into two categories. The first type is the Siamese model, which uses image pairs [11] or triplets [17] as inputs. In [18], Chen et al. proposed using quadruplet loss. Compared with the triplet loss, quadruplet loss can cause the model to have a smaller intra-class variation and a larger inter-class variation. A major shortcoming of the Siamese model is that it does not effectively employ pedestrian identity annotations. The second common type of CNN model is the classification model, which is often used in object detection [19] and image classification tasks [14]. Compared with the Siamese model, the classification model makes full use of pedestrian identity annotations. In [20], Zheng et al. regards each pedestrian as a category of the classification problem and uses the pedestrian identity as the label of the training data to train the CNN network. Such loss in this network is called identification loss. This kind of network is called an ID-discriminative embedding (IDE) network. The IDE network is a very important baseline benchmark in the area of person re-ID.

However, these deep-learning-based re-ID models heavily rely on the classification CNN backbones, such as ResNet [3], VGG [21], and SENet [4]. These CNN backbones were specifically designed to address the classification task, and experiments have been carried on the classification datasets. Due to the difference between the classification task and person re-ID datasets, the performance may be limited when using such classification-oriented networks to perform re-ID tasks [5]. To solve these problems mentioned above, we propose a retrieval network suitable for person re-ID tasks by constructing hierarchical residual-like connections within a single residual block, which we named the Re-Net network.

C. RES2NET MODULE

In many deep learning studies, deep convolutional features have proved to provide useful representations for classification or retrieval tasks. In person re-ID task, it is important to represent the multi-scale features of pedestrians. However, most of the existing methods represent multi-scale pedestrian features in a layer-wise manner. Gao et al. [22] proposed a novel Res2Net block for CNN, which is implemented by constructing hierarchical residual-like connections within a single residual block. Figure 1 compares the difference between the bottleneck block of ResNet50 and the Res2Net module.

As shown in Figure 1 (b), the Res2Net module replaces the $3 \times 3$ filters with 3 smaller groups of filters, while connecting these filter groups in a hierarchical residual-like style. Specifically, the feature maps from the previous layer first pass through a $1 \times 1$ convolution layer, and then the obtained feature maps are evenly split into 4 sub-feature maps, and each sub-feature map is defined as $X_i$, where $i \in \{1, 2, \ldots, 4\}$. The spatial size of each sub-feature map $X_i$ is the same, and the number of channels is 1/4 of the input feature maps. Except for $X_1$, all other $X_i$ have a corresponding $3 \times 3$ convolution layer, which is denoted by $K_i()$. The output of $K_i()$ is defined as $Y_i$. The sub-feature map $X_i$ is added with the output of $K_{i-1}()$, and then fed into $K_i()$, repeat this process until all input feature maps are processed. Thus $Y_i$ can be formulated as:

$$Y_i = \begin{cases} 
X_i & i = 1; \\
K_i(X_i) & i = 2; \\
K_i(X_i + Y_{i-1}) & 2 < i \leq 4; 
\end{cases}$$

Finally, all feature maps $Y_i$ are concatenated and sent to another $1 \times 1$ convolution layer to fuse features together.

III. PROPOSED METHOD

In this section, we will first describe the structure of the proposed Re-Net network in Section A, and then introduce two loss functions for the network in Section B. Finally,
Z. Cao, H. J. Lee: Learning Multi-Scale Features and Batch-Normalized Global Features for Person re-ID

FIGURE 2. The framework of our proposed network. Given a triple of images resized to $384 \times 128$ as input, the input images first pass through the FE-Net to acquire the global feature descriptors $f_t$, and then we let the features $f_t$ pass through the FN-Net to acquire batch-normalized global features $f_i$. In the training phase, the triplet loss and identification loss are calculated by using multi-scale global features $f_t$ and batch-normalized global features $f_i$ respectively. In the testing phase, we use the batch-normalized global features $f_i$ of pedestrians as feature descriptors to do person re-identification tasks.

we describe the method of measuring the distance between two features in Section C.

A. NETWORK ARCHITECTURE

For person re-ID tasks, the target is to retrieve and output a set of matching images from the gallery for a given query image. In this paper, we develop a Re-Net model based on a multi-scale backbone architecture. The proposed Re-Net model is supervised by identification loss and triplet loss together. The proposed Re-Net takes triplet images as mini-batch inputs and learns a discriminative embedding, such that the representations of the same-label image pairs are closer to each other, while the image pairs of the different labels are distant from each other. Figure 2 briefly illustrates the pipeline of the proposed Re-Net network.

The proposed Re-Net network consists of a FE-Net network, a FN-Net network, and an output module. The network takes triplet images as inputs, the inputs images first pass through the FE-Net to acquire the global feature descriptors $f_t$, and then inputs the features $f_t$ to the FN-Net to acquire batch-normalized global features $f_i$. The proposed network has multiple outputs and uses different features for the training and test stage. In the training phase, we use the global features $f_t$ to calculate the triplet loss, while the batch-normalized global features $f_i$ are used to calculate the identification loss. In the inference phase, we choose the batch-normalized global feature $f_i$ as the pedestrian descriptor that is used to perform the person re-ID task.

B. FEATURE EXTRACTION NETWORK

For person re-ID task, pedestrian accessories may appear in the image with different sizes, such as the backpack and pair of shoes in Figure 3 (c). In addition, the necessary contextual information of pedestrian accessories may occupy a larger area in the image than itself. For example, in Figure 3 (b), we need to rely on the hand and shopping bag as context to determine whether the black object in the hand is a mobile phone or an umbrella. Therefore, it is important to represent features at multiple scales.

To obtain multi-scale representations of pedestrian features in person re-ID tasks, feature extractors need to adopt a wide range of receptive fields to describe pedestrians and their accessories at different scales. At present, the most commonly used method is to represent multi-scale features of pedestrians in a layer-by-layer manner. Gao et al. [22] proposed a novel Res2Net block for CNN, which is implemented by constructing hierarchical residual-like connections within a single residual block. The Res2Net module represents multi-scale features at the granularity level and increases the receiving field range of each network layer. In addition, as a
FIGURE 3. Multi-scale representations are essential for understanding pedestrians and their accessories. (a) to (c): some examples of objects appearing with different sizes in a single image.

FIGURE 4. Comparison of the Res2Net module and proposed IB-Res2Net module. (a) Res2Net module; (b) proposed IB-Res2Net module.

cross camera retrieval task, the variation of the pedestrian image style has a great impact on the performance of the person re-ID model. The pedestrian image captured by different cameras is the main reason for such variations.

To solve the problems mentioned above, we propose an improved Res2Net module, named IB-Res2Net module, which improves the representation ability of pedestrian multi-scale features at a more fine-grained level. Furthermore, motivated by the IBN-Net [23], to improve generalizable feature learning, we introduce instance normalization (IN) layers into the IB-Res2Net module to cope with cross-camera image discrepancies. The specific method is shown in Figure 4(b). After the first $1 \times 1$ convolution layer of the residual block, we adopt a batch normalization (BN) layer for half of the channels and apply the IN layer for the remaining channels. With the IN layer and BN layer, it is possible to greatly enhance the residual module’s ability in one domain and generalization ability in another domain without fine-tuning.

In our proposed IB-Res2Net module, splits are processed in a multi-scale fashion, which improves the ability of the CNN model to the extraction of both pedestrian and their accessories fine-grained feature information. To better fuse features extracted by the five groups of $3 \times 3$ filters at different scales, we connect all the sub-feature maps and output them through a $1 \times 1$ convolution layer. By using split and concatenation strategies, the residual module can be forced to process features more efficiently.

The structure of Re-Net before the adaptive average pooling (AAP) layer remains the same as the Res2Net [22] network. The difference from the original Res2Net is that the pooling layer and following layers are removed.

### C. FEATURE NORMALIZATION NETWORK

To obtain better performance, a popular approach is to train the re-ID model by combining identification loss and triplet loss. However, most of the earlier related works combined these two losses to constrain the same pedestrian features. Identification loss aims to construct several hyperplanes and divide the features of different classes into different subspaces. Therefore, in the inference phase, for the re-ID model optimized by identification loss, using cosine distance as the distance metric function will obtain better feature measurement. On the other hand, the triplet loss is calculated by the Euclidean distance, which improves the intra-class compactness and the inter-class separability of pedestrian features in Euclidean space. Therefore, the targets of triplet loss are inconsistent with the targets of identification loss in the embedding space. For pedestrian image pairs distributed in the embedding space, identification loss is mainly used to optimize their cosine distances, while triplet loss is mainly concentrated on the Euclidean distances. In this case, if we use these two losses to simultaneously optimize a feature vector, identification loss will reduce the intra-class compactness of triplet loss, and triplet loss will affect the clear decision surface of identification loss. In the training phase, there will be a phenomenon in which one loss is decreasing, while the other is oscillating or even increasing.

Xiong et al. [24] propose to add a batch normalization layer between the pooling layer and fully connected (FC) layer to overcome overfitting and boost the performance of re-ID models. Similarly, in [25] the author claims that adding a batch normalization layer before the FC layer can make the distribution of features embedded in space smoother.

Motivated by [24], [25], we propose an FN-Net to prevent overfitting and overcome the problem of inconsistency between the targets of two losses in the embedding space. The structure of the FN-Net is shown in Figure 2. The FN-Net is composed of a linear layer, a batch normalization layer, and a dropout layer. Dropout randomly discards the output of each hidden neuron with a probability during the training phase. The FN-Net is added after the AAP layer and before the FC layer. The global features before the FN-Net are denoted as $f_t$. Then, we let the global features $f_t$ pass through the FN-Net to acquire the batch-normalized global features $f_i$.

In the training stage, the global features $f_t$ and the batch-normalized global features $f_i$ are used to calculate the triplet loss.
loss and identification loss, respectively. The experimental results show that the FN-Net significantly improves the performance of the re-ID models.

D. LOSS FUNCTION
1) IDENTIFICATION LOSS
The softmax function used in the classification task can be formulated by the following:

\[ \hat{p}_i = \frac{\exp(a_i)}{\sum_{i=1}^{n} \exp(a_i)}, \]  

(2)

where \( \exp(x) \) represents the exponential function of \( e^x \) (\( e \) is the Napier constant \( 2.71828 \ldots \)). \( \hat{p}_i \) indicates the output of predicted probability, \( n \) indicates the number of samples. In this work, we use softmax with cross-entropy loss \( (L_s) \) as the identification loss, which is computed as:

\[ L_s = -\sum_{i=1}^{n} p_i \log(\hat{p_i}), \]

(3)

where \( \hat{p}_i \) indicates the output of predicted probability, \( p_i \) indicates the target probability, and \( n \) indicates the number of samples.

2) TRIPLET LOSS
The triplet loss aims to learn discriminative representations, which makes the distance between the anchor image and positive image in the embedding space small and makes the distance between the anchor image and negative image large. A triplet example in our work consists of an anchor image \( a \), a positive image \( p \) (with the same identity as the anchor), and a negative image \( n \) (different from the identity of the anchor). Then, the triplet loss \( (L_t) \) function used in the Re-Net model can be formulated as:

\[ L_t = \sum_{i=1}^{n} \max(d(a_i, p_i) - d(a_i, n_i) + \text{margin}, 0), \]

(4)

where \( a, p, n \) indicate the anchor, positive, and negative examples, respectively. \( d(a_i, p_i) \) indicates the feature distances of a positive pair, \( d(a_i, n_i) \) indicates the feature distances of a negative pair. Moreover, to promote the discriminative capability of the Re-Net network, we add a constant margin to produce a gap between \( a \) and \( p \) versus \( a \) and \( n \), as shown in Equation (4). The margin indicates the margin of triplet loss. In this work, the margin is set to 0.3.

3) TOTAL LOSS
To align with the person re-ID problem and leverage the mutual benefits of triplet loss and identification loss, in this work, we consider a mixture of \( L_s \) and \( L_t \) as follows:

\[ Loss = \beta L_s + (1 - \beta) L_t, \]

(5)

where \( \beta \) is the balanced weight of the two losses. In section V, we evaluated the impact of different weight \( \beta \) on the performance of the model. We observed that when \( \beta \) is set to 0.6, the model can achieve the best performance. More experimental results can be obtained in Section V.

IV. EXPERIMENTS
In this section, we empirically evaluate the proposed Re-Net model. We introduce the benchmark datasets in Section A and the Evaluation Protocol in Section B. The implementation details are described in Section C.

A. DATASETS
1) MARKET-1501 DATASET
The Market-1501 dataset [26] was collected in front of a supermarket on the campus of Tsinghua University. The whole dataset was captured by 6 cameras, which included a low-resolution camera and five high-resolution cameras. This dataset contains \( 32,668 \) labeled bounding boxes of 1,501 identities. In this work, we divided the dataset into four groups, with 12,185 images with 751 identities used as the training set, 751 images with 751 identities used as the validation set, 3,368 images with a different set of 750 identities used as the query set, and 19,732 images with 750 identities used as the gallery set.

2) DukeMTMC-reID DATASET
The DukeMTMC-reID dataset [27] is a subset of the DukeMTMC dataset for image-based person re-ID. The DukeMTMC is a dataset of surveillance camera footage collected at Duke University. The whole dataset contains \( 1,104 \) identities that appear in more than two cameras, and \( 408 \) identities (distractor ID) that only appear in only one camera, \( 2,228 \) queries, \( 17,661 \) gallery images, and \( 16,522 \) training images. In this work, we divided the dataset into four groups, with \( 15,820 \) images with 702 identities as the training set, \( 702 \) images with 702 identities as the validation set, \( 2,228 \) images with a different set of 702 identities as the query set, and \( 17,661 \) images with \( 1,110 \) identities (702 ID + 408 distractor ID) as the gallery set.

3) CUHK03 DATASET
The CUHK03 dataset [28] was collected from The Chinese University of Hong Kong (CUHK) campus. The whole dataset contains \( 28,192 \) bounding boxes and 1,467 identities. Among them, each identity is captured by two cameras on the campus of CUHK, each camera has an average of 4.8 images. The dataset provides both DPM-detected bounding boxes (detected set) and manually labeled bounding boxes (labeled set), and we use the former in this paper. We follow the new training/testing protocol proposed by Zhong et al. [29]. In this paper, we divided the dataset (detected set) into four groups, with \( 6,598 \) images with 767 identities as the training set, \( 767 \) images with 767 identities as the validation set, \( 1,400 \) images with the other 700 identities as the query set, and \( 5,332 \) images with 700 identities as the gallery set.

For the Market-1501 and DukeMTMC-reID datasets used in this work, we adopted the evaluation packages provided by Zheng et al. [26] and Zheng et al. [30], respectively. For the CUHK03 dataset, we adopted the new training/testing protocol proposed by Zhong et al. [29]. All the experiments
TABLE 1. The details of datasets used in our experiments.

| Datasets   | Market1501 | Duke MTMC-reID | CUHK03 |
|------------|------------|----------------|--------|
| Train      | 751        | 702            | 767    |
|            | 12,185     | 15,820         | 6,598  |
| Validation | 751        | 702            | 767    |
|            | 751        | 702            | 767    |
| Query      | 750        | 702            | 700    |
|            | 3,368      | 2,228          | 1,400  |
| Gallery    | 750        | 1,110          | 700    |
|            | 19,732     | 17,661         | 5,332  |

evaluate the single-query setting. Some example images from these three datasets are shown in Figure 5. The detailed information of the datasets used in our experiments is shown in Table 1.

B. EVALUATION PROTOCOL

The mean average precision (mAP) and cumulative matching characteristics (CMC) curve are the two most popular evaluation metrics for person re-ID tasks.

1) DISTANCE METRIC

There are many commonly used distance metric functions, such as the Euclidean distance, Manhattan distance, cosine distance, and Hamming distance. Under different distance metric functions, the performance of the model retrieval process is not the same. A good distance metric function is helpful to improve the performance of the model.

In this work, the similarity of two vectors is computed using the cosine distance. If there are two input images $i_1$ and $i_2$, we extract visual features by using the Re-Net model and obtain two feature vectors, denoting $A$ and $B$ respectively. Then, the distance between the two images in the feature space can be defined as:

$$\text{Similarity} = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}.$$  \hspace{1cm} (6)

where $i$ denotes the image index in the dataset.

2) mAP

mAP is used to evaluate the overall performance of the re-ID model. The mAP is the mean value of the average precision of all queries. The average precision is defined as the area under the precision-recall curve. That is, if the gallery set of relevant images for a query image is $q_j \in Q$ is $\{d_1, \ldots, d_m\}$ and $R_{jk}$ is the set of ranked retrieval results from the top result until you get image $d_k$. Then, the mAP can be formulated as:

$$\text{mAP} (Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} P (R_{jk}),$$  \hspace{1cm} (7)

where a relevant gallery image is not retrieved at all, the precision value in the above equation is taken to be 0.

Because the mAP not only considers the precision of the algorithm but also considers the recall of the algorithm, it provides a more comprehensive evaluation.

3) CMC

The CMC curve is used to assess the accuracy of algorithms that produce an ordered list of possible matches. CMC evaluates the top $n$ nearest images in the gallery set with respect to one query image. For example, if the correct match of a query image is at the $k^{th}$ position ($k \leq n$), then this query is considered as a success of rank $n$. The final CMC curve is computed by averaging the accuracy of the CMC top-k over all of the queries.

To evaluate the performance of our proposed Re-Net model in practice and compare its results with other state-of-the-art re-ID methods, in this paper, the accuracy at rank-1, rank-5, and rank-10 and the mAP are reported as evaluation metrics.

C. IMPLEMENTATION DETAILS

In this experiment, we implemented our proposed method using the PyTorch deep learning framework. The whole experiment is divided into three stages, which are the image preprocessing, training, and testing stages. The following is a detailed description of each stage.

1) IMAGE PREPROCESSING

In person re-ID, pedestrian images are sometimes occluded by other objects or pedestrians. To overcome the occlusion problem and improve the generalization ability of our model, before the image input network training, we used a random erasing augmentation (REA) method to preprocess pedestrian images. The REA augmentation method was proposed by Zhong et al. [31] in 2017. In practice, we randomly select a rectangle region in the image and erase its pixels with probability (p) of performing the random erasing operation to 0.5. Some examples are shown in Figure 6.
images are augmented by a horizontal flip, normalization, REA, and resized to $384 \times 128$.

2) TRAINING PHASE
The mini-batch of triplet loss includes $B = P \times K$ images, where $P$ and $K$ denote the number of different identities and the instances of different images per identity, respectively. In practice, we randomly sample 6 identities each with 8 instances and set the mini-batch size to 48. The training images are augmented by horizontal flip, normalization, REA [31], and resized to $384 \times 128$. The proposed FE-Net model is pre-trained on ImageNet [32], while the FN-Net is initialized using the Kaiming initialization method [33]. We train the proposed Re-Net model for 120 epochs, using Adam as the optimizer, and we utilized the multi-step learning rate schedule. The initial learning rate is set to 0.00035, and it decays by 0.1 at the 40-th and 70-th epochs. Since there are an identification loss function and a triplet loss function in our proposed network, we first calculate all gradients produced by each loss function and then add the gradients together to update the network. In this paper, the balance weight $\beta$ of the two losses is set to 0.6. The validation accuracy and loss curves of the proposed model on the three datasets are shown in Figure 7 and Figure 8.

It can be observed form the validation accuracy curves and validation loss curves that the model achieves the peak accuracy after 70 epochs.

3) TESTING PHASE
In the testing stage, we choose $f_i$ as the pedestrian descriptors to perform the person re-ID task. Given an image resized to $384 \times 128$ as input, we feed forward this image to the Re-Net model and obtain a 1024-dim pedestrian descriptor $f_i$. These descriptors are stored offline when the descriptor of all the images in the gallery set is obtained. When a query input image is given, the network will extract its descriptor online; then sort the cosine distance between the query image and all the gallery image features to obtain the final ranking result. The flow chart of the testing phase is shown in Figure 9.

V. RESULTS AND ANALYSIS
In this section, we evaluate our proposed method on three large-scale person re-ID benchmark datasets and show the results of the Re-Net model compared with other state-of-the-art methods. The Rank-1 accuracy and mAP are reported as evaluation metrics. In addition, to evaluate the effect of each component in our proposed Re-Net model, we perform extensive ablation studies on the DukeMTMC-reID datasets and Market1501 datasets. We show the results in Tables 2, 3, and 4.
A. ANALYSIS OF RE-NET MODEL
The IDE model specified in [1] is usually used as the baseline for person re-ID tasks [34]. For comparison, we implemented the IDE model using Res2Net and FE-Net as backbone networks, denoted as baseline-R and baseline-F, respectively. Both baseline-R and baseline-F are trained with neither the FN-Net nor the triplet loss. Then, we sequentially added the FN-Net and triplet loss based on baseline-F and evaluated the performance of the model. The experiential results are shown in Table 2. For simplicity, we do not use $k$-reciprocal re-ranking algorithms [29], which considerably improves the mAP. The experiential results show that our proposed FE-Net and FN-Net can significantly improve the performance of the re-ID model.

B. ANALYSIS OF TWO LOSSES
The proposed Re-Net model is supervised by two losses. In our work, we set a weight $\beta$ to balance two losses, as shown in Equation (5). In this section, we evaluate the impact on model performance by using different weight $\beta$. We evaluate the performance on two large-scale benchmark datasets and show the results in Table 3. For simplicity, we do not use $k$-reciprocal re-ranking algorithms [29] in this study.

We observed that different weights have a large impact on the performance of the Re-Net model. When the weight $\beta$ is set to 0.6, the Re-Net model obtains the best performance.

C. COMPARISON WITH STATE-OF-THE-ART RE-ID METHODS
We evaluate our proposed re-ID model compared with lots of existing state-of-the-arts methods on three large-scale benchmark datasets. The compared methods have been divided into three groups, i.e., deep learning methods that are pose-guided, deep learning methods with part features and deep learning methods with the global feature. All the experiments evaluate the single-query setting. The results are shown in Tables 4 and 5. RK indicates that the k-reciprocal re-ranking algorithm is used.

1) EVALUATIONS ON THE DukeMTMC-reID DATASET
We compare the proposed Re-Net model with eleven latest methods on DukeMTMC-reID datasets. As shown in Table 4, with the re-ranking scheme [29], our model achieves a rank-1 accuracy of 92.11% and a mAP of 90.10%, and it outperforms other state-of-the-art re-ID algorithms on this dataset.
TABLE 4. Comparison with the state-of-the-art methods on the Market-1501 and DukeMTMC-reID datasets. RK indicates that the k-reciprocal re-ranking algorithm is used.

| Type               | Methods           | Market1501 Rank=1 | Market1501 Rank=5 | Market1501 Rank=10 | DukeMTMC-reID Rank=1 | DukeMTMC-reID Rank=5 | DukeMTMC-reID Rank=10 | mAP |
|--------------------|-------------------|--------------------|--------------------|--------------------|-----------------------|-----------------------|-----------------------|-----|
| Pose-guided        | PSE [37]          | 87.70              | -                  | -                  | 69.00                 | 79.80                 | -                     | -   |
|                    | PSE (RK) [37]     | 90.30              | -                  | -                  | 84.00                 | 85.20                 | -                     | 79.80|
|                    | PIE [35]          | 87.33              | 95.56              | 96.81              | 69.25                 | 80.84                 | 88.30                 | 90.72|
|                    | GLAD [38]         | 89.90              | -                  | -                  | 73.90                 | -                     | -                     | -   |
| Part feature       | SCPNet [39]       | 94.10              | 97.70              | -                  | 81.80                 | 84.80                 | 91.90                 | -   |
|                    | AlignedReID [40]  | 91.80              | -                  | -                  | 79.30                 | -                     | -                     | -   |
|                    | AlignedReID (RK)  | 94.40              | -                  | -                  | 90.70                 | -                     | -                     | -   |
|                    | PCB [50]          | 93.80              | 97.50              | 98.50              | 81.60                 | 83.30                 | 90.50                 | 92.50|
|                    | BFE [41]          | 95.30              | -                  | -                  | 86.70                 | 89.00                 | -                     | 76.00|
|                    | Pyramid [42]      | 95.70              | 98.40              | 99.00              | 88.20                 | 89.00                 | -                     | 79.00|
|                    | PCN+PSP [43]      | 92.81              | 97.24              | 98.13              | 78.80                 | 85.37                 | 92.77                 | 94.52|
|                    | (RK)              | 94.39              | 96.88              | 97.51              | 90.79                 | 89.41                 | 93.54                 | 95.15|
| Global feature     | IDE [20]          | 79.51              | -                  | -                  | 59.87                 | -                     | -                     | -   |
|                    | SVTLDet [36]      | 82.30              | 92.30              | 95.20              | 62.10                 | 76.70                 | 86.40                 | 89.90|
|                    | AWTL [44]         | 89.46              | -                  | -                  | 75.67                 | 79.80                 | -                     | 63.40|
|                    | TriNet [45]       | 84.90              | 94.21              | -                  | 69.14                 | -                     | -                     | -   |
|                    | TriNet (RK) [45]  | 86.67              | 93.38              | -                  | 81.07                 | -                     | -                     | -   |
|                    | BoT [25]          | 94.50              | -                  | -                  | 85.90                 | 86.40                 | -                     | 76.40|
|                    | (RK)              | 95.40              | -                  | -                  | 94.20                 | 90.30                 | -                     | 89.10|
|                    | Ours              | 94.86              | 98.49              | 99.13              | 87.31                 | 89.13                 | 94.47                 | 96.54|
|                    | (RK)              | 96.03              | 98.02              | 98.37              | 94.80                 | 92.11                 | 95.24                 | 96.63|

2) EVALUATIONS ON THE MARKET-1501 DATASET
We compare the proposed Re-Net model with other state-of-the-art re-ID models on Market-1501 datasets. We obtain a rank-1 accuracy of 96.03% and a mAP of 94.80%. As shown in Table 4, the Re-Net model outperforms other state-of-the-art re-ID models.

3) EVALUATIONS ON THE CUHK03 DATASET
The CUHK03 dataset provided two types of pedestrian bounding boxes: manually labeled (labeled set) and DPM-detected (detected set), and we use the latter in this paper. As shown in Table 5, we compare our Re-Net model with nine state-of-the-art re-ID methods on the CUHK03 datasets. Our proposed Re-Net model obtains a rank-1 accuracy of 84.14% and a mAP of 84.62%, it outperforms other state-of-the-art re-ID methods.

TABLE 5. Comparison with the state-of-the-art methods on CUHK03-datasets. RK indicates that the k-reciprocal re-ranking algorithm is used.

| Method            | CUHK03-Detected Rank=1 | mAP |
|-------------------|-------------------------|-----|
| PAN [46]          | 36.30                   | 34.00 |
| SVTLDet [36]      | 41.50                   | 37.30 |
| HA-CNNF [47]      | 41.70                   | 38.60 |
| AOS [48]          | 47.70                   | 43.40 |
| MLFND [49]        | 52.80                   | 47.80 |
| PCB [50]          | 63.70                   | 57.50 |
| Mano [51]         | 65.50                   | 60.50 |
| DGN [52]          | 65.60                   | 61.10 |
| PCN+PSP [43]      | 60.71                   | 56.01 |
| Ours              | 75.71                   | 71.31 |
| (RK)              | 84.14                   | 84.62 |

D. INSTANCE RETRIEVAL RESULTS
In this section, we apply the proposed Re-Net network to the task of generic pedestrian retrieval. We visualized some retrieval results on three benchmark datasets, as shown in Figures 10, 11, and 12. In these figures, the leftmost image of each row is the query image, which is shown in a black box. The images on the right-side are the retrieval images sorted from left to right according to the similarity score with the query image and are denoted by 1-10, respectively. Among
them, the label of the correctly matched image is drawn in green boxes and the labels of the incorrectly matched images are drawn in red boxes. It is observed from the retrieving samples that the proposed model is also robust to partial pedestrian occlusion.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a new retrieval architecture to solve the task of person re-ID, which we named the Re-Net network. Being different from many state-of-the-art methods that use the complex networks, the proposed retrieval backbone network is implemented using only global features and simple triplet loss and identification loss. In the backbone network is implemented using only global features and simple triplet loss and identification loss. In the Re-Net Bottleneck block, we replace a group of 3 × 3 filters with groups of 3 × 3 filters with 5 groups of 3 × 3 filters, while connecting different filter groups in a hierarchical residual-like style. The experimental results show that our proposed network improves the representation ability of pedestrian multi-scale features at a more fine-grained level. Furthermore, we proposed an FN-Net to overcome the problem that the two types of losses are inconsistent in the embedding space. The Re-Net model outperforms the state-of-the-art methods on three popular person re-ID benchmarks datasets and shows a stronger fine-grained pedestal representation ability for application of the instance retrieval task.

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] E. Ahmed, M. Jones, and T. K. Marks, “A Siamese long short-term memory architecture for human re-identification,” in Proc. IEEE Comput. Vis. Pattern Recognit., Jun. 2015, pp. 3908–3916.
[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[4] L. Zhang, X. Chen, J. Zhang, and X. Huang, “Beyond triplet loss: A deep quadruplet network for person re-identification,” in Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 403–412.
[5] G. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.
[6] R. Zhong, L. Zheng, and Y. Yang, “A discriminatively learned CNN embedding for person reidentification,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 14, no. 1, pp. 1–20, Jan. 2018.
[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556
[8] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. S. Torr, “Res2Net: A new multi-scale backbone architecture,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Aug. 30, 2019, doi: 10.1109/TPAMI.2019.293857.
[9] X. Pan, P. Luo, J. Shi, and X. Tang, “Two at once: Enhancing learning and generalization capacities via ibn-net,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 464–479.
[10] F. Xiong, Y. Xiao, Z. Cao, K. Gong, Z. Fang, and J. T. Zhou, “Towards good practices on building effective CNN baseline model for person re-identification,” 2018, arXiv:1807.11042. [Online]. Available: http://arxiv.org/abs/1807.11042
[11] H. Luo, Y. Gu, X. Liao, S. Bai, and W. Jiang, “Bag of tricks and a strong baseline for deep person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2019.
[12] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1116–1124.
[13] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-task, multi-camera tracking,” in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer, 2016, pp. 17–35.
[14] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Pattern Recognit., Jun. 2014, pp. 152–159.
[15] 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., Jul. 2017, pp. 1318–1327.
[16] Z. Zhong, L. Zheng, and Y. Yang, “Unlabeled samples generated by GAN improve the person re-identification baseline in vitro,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 3754–3762.
[17] 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
[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.
[19] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1026–1034.
[20] T. Xiao, H. Li, W. Ouyang, and X. Wang, “Learning deep feature representations with domain guided dropout for person re-identification,” in Proc. IEEE Comput. Vis. Pattern Recognit., Jun. 2016, pp. 1249–1258.
[21] 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.
[22] Y. Sun, L. Zheng, W. Deng, and S. Wang, “Svndet for pedestrian retrieval,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 3800–3808.
[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., Jun. 2018, pp. 420–429.
[38] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, “GLAD: Global-local-alignment descriptor for pedestrian retrieval,” in Proc. ACM Multimedia Conf. (MM), 2017, pp. 420–428.

[39] X. Fan, H. Luo, X. Zhang, L. He, C. Zhang, and W. Jiang, “Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification,” in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2018, pp. 19–34.

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

[41] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, “Batch DropBlock network for person re-identification and beyond,” 2018, arXiv:1811.07130. [Online]. Available: http://arxiv.org/abs/1811.07130

[42] 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.

[43] P. Chikontwe and H. J. Lee, “Deep multi-task network for learning person identity and attributes,” IEEE Access, vol. 6, pp. 60801–60811, 2018.

[44] E. Ristani and C. Tomasi, “Features for multi-target multi-camera tracking and re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6036–6046.

[45] 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

[46] 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.

[47] W. Li, X. Zhu, and S. Gong, “Harmonious attention network for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2285–2294.

[48] 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., Jun. 2018, pp. 5098–5107.

[49] 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., Jun. 2018, pp. 2109–2118.

[50] 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., 2018, pp. 480–496.

[51] 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 Proc. Eur. Conf. Comput. Vis., 2018, pp. 365–381.

[52] 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.

---

ZONGJING CAO is currently pursuing the master’s degree in computer science and engineering with Jeonbuk National University, Jeonju, South Korea. His research interests include deep learning, computer vision, image processing, image retrieval, and person re-identification.

HYO JONG LEE (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from The University of Utah, USA. He was involved in computer graphics, image processing, and parallel processing. He is currently a Professor of the Division of Computer Science and Engineering and the Director of the Center for Advanced Image and Information Technology, Jeonbuk National University, Jeonju, South Korea. His research interests include image processing, artificial intelligence, medical imaging, parallel algorithms, and brain science.