Computer Science and Information Systems 21(4):1483–1498 https://doi.org/10.2298/CSIS230829044L

# Learning Discriminative Representations through an Attention Mechanism for Image-based Person Re-identification

Jing Liu[1] and Guoqing Zhou[2]

[1] School of Computer Science, Weinan Normal University,
Weinan 714099, Shaanxi, China
liujing8318@mail.nwpu.edu.cn
[2] School of Computer Science, Northwestern Polytechnical University,
Xi'an, 710072, Shaanxi, China
zhouguoqing@nwpu.edu.cn

**Abstract.** Over the past years, person re-identification has been obtaining various attentions in computer vision tasks. However, existing methods mainly focus on building massive number of deep architecture layers, which is unsuitable for extracting the robust features for person re-ID. In this paper, we present a novel hybrid framework PGAN, through which the discriminative representations can be learned for person re-ID. Specifically, a novel self-attention method named channel-wise attention mechanism is adopted to learn the informative representations from the patch-network and global network, respectively. In addition, CSwin Transformer is exploited to re-extract the discriminative features from the residual blocks. We obtain a mAP of 81.8% and 80.3% of the labeled and detected dataset on the CUHK0-NP dataset. And we obtain a mAP of 83.4% and 91.3% on the DukeMTMC and Market-1501 datasets respectively. Comprehensive experiments are performed on the three datasets, (Market-1501, DukeMTMC-reID and CUHK03-NP), demonstrating the efficiency of the introduced approach.

**Keywords:** person re-identification, channel-wise attention, deep learning.

## 1. Introduction

Person re-identification (re-ID) task aim to identify a person through video samples from different direction cameras. Along with the rapid development of deep learning (DL), various researchers have shifted from hand-crafted feature engineering [26] to deep learning techniques [19,39,50,54,44,46,45,40,12]. For example, in [53,21,51], the authors adopt DL technology to capture the robust features from the video samples for person re-ID. In addition, the attention mechanism attempts to be adopted in numerous studies [23,20,18,23,15,49,35] to represent the notable parts (e.g., spatial locations and etc.) by convolutional responses and filter the useless parts (e.g., background). Now, different attention schemes are adopted to represent great performances for person re-ID [20]. Yet, the general attention approaches can only be adopted to capture the coarse features, which is not able to learn complicated features from the visual sections for person re-ID. Moreover, several features can include some informative representations of attentional feature maps for person re-ID. Hence, the authors use a new attention mechanism to learn high-level representations. In addition, to extract the robust representations with all instances,


**Fig. 1.** These images sampled from four person re-ID datasets, representing that informative representations are important for person re-ID. Each sub-figure group consists of the three sub-figures (query, true, false)

the researchers consider that all data instances can be seen as a a similarity graph in person re-ID task. Also they combine the high-level and spectral features to promote the accuracy for person re-ID. Especially, through [27,56,41,48], the important features of person re-ID are also learned. And [27] proposes a deep graph metric learning (DGML) for person re-ID, which can learn the consistency from intra-video spatial graphs of consecutive frames, then the spatial graph learns neighborhood relationships of the detected person instances of each frame. For the another aspect, DGML can capture the distinguished features among the inter-video spatial graphs for person re-ID. In addition, the development of hand-crafted features needs a great number of factors (e.g., domain experience, time, etc.). Hence, bags and books can result in the poor performances of person re-ID. Furthermore, to learn the features of bags, scientists should learn domain-specific knowledge, which needs lots of time. This is the motivation of our work to design a robust architecture for person re-ID.

Up to now, through the various end-to-end DL methods, promising performances has been achieved in person re-ID, and the following problems still existed. The first problem is the representation of discriminative features. Current methods are mainly focused into two aspects by the point of feature representations (i.e., hand-crafted and deep learned features.). Though hand-crafted features have been proven to be obtaining in person re-ID based on hand-crafted features. Yet, there are still exists several problems. For instance, the designed of hand-crafted features acquire numerous aspects (e.g., domain experience, labor, etc). For example, bags and books can influence the accuracy for person re-ID. In addition, to capture the features of bags, researchers need to acquire task-specific knowledge to waste a lot of time. As shown in Fig. 1, to capture the informinative representations in the images, one should overcome this difficulties (such as the black bag in (a)) in the databases. For the rest of the examples in Fig. 1, we can observe that the bags can bring some problems to feature learning. More importantly, different databases have the same problems while extracting the informative features, in which a robust architecture also needs to be designed for feature learning.

In addition, another problem is the attentive feature learning. Although most of the DL methods for person re-ID have a discriminative feature extraction ability, the importance of the association between high-level semantic feature features and channels is ignored in

the feature learning step in person re-ID. For example, as shown in Fig. 1 (a), to extract the informative features, we should filter the bags of noisy features and capture the discriminative features for person re-ID. Also, the useless features lead to a reduced scalability of the model, which makes it hard for real-world large-scale usage. Meanwhile, the useless features can result in the problem of overfitting in the process of model training, which also brings some barriers for the real-world application.

In our work, to address the above-mentioned issues, we propose a novel architecture, named patch-global attention network (PGAN), for addressing person re-ID. Specifically, PGAN contains two components, i.e., patch-network and global-network, which also contains a channel-wise attention mechanism to learn the patch-global feature representations. For the patch-network, our aim is to learn the patch features, e.g., the face, the body, and other patches. This is because that different parts of the body contain different semantic information of features. Therefore, these different convolutional neural networks (CNN) branches can be encoded in the patch-network and informative feature representations for person re-ID can be captured. Through the global-network the global representations for person re-ID can be learned. More importantly, a channel-wise attention mechanism is adopted to learn the local and global features as well as fuse the different scale features for person re-ID. In addition, Swin Transformer is used to re-extract the discriminative features from the residual blocks.

The contributions of the presented paper are three-aspects:

- We introduce a novel framework, named PGAN, for modeling the multi-scale features for person re-ID. CSwin Transformer is adopted to extract the informative features from the residual blocks.
- In the architecture, a novel attention method, named channel-wise attention mechanism is adopted to learn the discriminative information through the patch-network and global-network, respectively.
- Comprehensive experiments are performed on some benchmark datasets, i.e., Market-1501, DukeMTMC-reID and CUHK03-NP, demonstrating the efficiency of the introduced approach.

The remainder of the presented paper is structured as follows. Several existing works are described in Section 2. In Section 3, the proposed deep architecture is described. The used databases and experimental performances are described in Section 4. In Section 5, conclusions and future works are described.

## 2. Related Works

In this part, we briefly overview the existing works on person re-ID. We first introduce the hand-crafted features, and describe the deep features for person re-ID.

### 2.1. Hand-crafted Features of Person Re-ID

Hand-crafted features are extracted for person re-ID in various studies [13]. A weighted color topology (WCT) feature is designed to model the spatial distribution features. In [47], it is found that the fusion of histogram features of oriented gradient (HOG) descriptor, HSV, and color histogram features is important for person re-ID. Meanwhile, in

[9], the authors also propose to use region-based feature salience for person re-ID. However, there still exists the problem that we have not extracted the high-level features, e.g., context information and semantic features. Hence, various studies are emerged with the development of DL. We briefly introduce the deep-learned features of person re-ID as follows.

### 2.2. Deep-learned Features in Person Re-ID

In the work of [28], Rao et al. propose a self-supervised gait encoding method through which unlabeled skeleton data can be learned to capture gait representations for person re-ID. The core idea of the introduced approach is that high-level semantics information is learned through a self-supervision mechanism. In addition, they also propose a locality-aware attention mechanism and a locality-aware contrastive learning scheme to model and maintain locality-awareness on an intra-sequence level and inter-sequence level in the self-supervised learning step. Finally, a new feature named constrastive attention-based gait encodings (CAGEs) is proposed to characterize gait effectively.

Bai et al. [1] collect a new large-scale dataset for person re-ID. Additionally, they also proposed a domain generalization re-ID architecture named dual-meta generalization network (DMG-Net), to learn the merits of meta-learning in the training step.

Two constructive modules are proposed in [2], i.e., a rectification domain-specific batch normalization (RDSBN) module and a graph convolutional network (GCN) based multi-domain information fusion (MDIF) module for the domain-specific view and domain-fusion view, respectively. RDSBN can be adopted to simultaneously reduce domain-specific characteristics and increase the distinctiveness of personal features. MDIF can be used to minimize field distances via combing representations of different fields.

In [42], the authors introduce a new pairwise loss function, through which fine-grained features can be learned by enforcing an exponential penalization on the images of small differences and a bounded penalization on the images of large differences.

Deep graph metric learning (DGML) is introduced of [27] for person re-ID, through which the consistency from intra-video spatial graphs of consecutive frames can be re-weighted, which the spatial graph capture neighborhood relationship in the detected person instances of each frame. From the another perspective, by using the DGML approach, the inter-video spatial graphs learned by numerous camera views at different aspects are identified. In order to further represent deploy a weak supervision into the DGML and solve the problem of weakly supervised person re-ID, weakly supervised regularization (WSR) is introduced to learn discriminative features by means of a weak identity loss and a cross-video alignment loss.

In [11], Hou et al. introduces a spatial-temporal architecture for video-based person re-ID. Concretely, a bilateral complementary network (BiCnet) is proposed to model spatial complementary information. The combination of a detail branch and context branch is used to model the original and long-range characteristic information. For every branch, biCnet can combine multiple parallel and diverse attention modules to discover divergent body parts for consecutive frames to obtain an integral characteristic of target identity. Moreover, a new block is designed to capture different scale features. The new block can be easily embed into BiCnet for spatial-temporal modeling.

Modeling static images with identity-related and unrelated features are proposed in [7]. Generative adversarial network is adopted to represent identity-related and unrelated

features of person images. Specifically, an identity-shuffling method is used to reduce the noisy labels without any additional methods.

It tries to overcome the unsupervised domain adaptation in [14] from a novel perspective for person re-ID. A cluster-wise contrastive learning algorithm (CCL) is adopted to maintain the feature extraction and to learn noise-tolerant features by a new unsupervised scheme. Meanwhile, they adopt a progressive domain adaptation (PDA) method is used to narrow the domain gap between source and target data. After that, fourier augmentation scheme is also utilized to maximize the size of class separability of re-ID models.

In [56], the authors propose a new architecture, named omni-scale network (OSNet) to model the different spatial scale features. The combination of multiple convolutional streams in the basic building block is used to learn the deep features. A unified aggregation gate of the omni-scale feature learning is adopted to add multi-scale features based on channel-wise weights. Meanwhile, the problem of cross-dataset discrepancies is also addressed on instance normalisation (IN) layers.

In [43], the authors introduce a new architecture based on GNN to overcome the above two studies on re-ID, i.e., group-aware person re-ID and group re-ID. Concretely, the authors formulate a context graph with group members as its nodes to consider dependencies among different people. Also, the authors develop a multi-level attention that can be formed in two group contexts, with an additional self-attention module for robust graph-level representations by attentively aggregating node-level features.

Based on the above-mentioned methods, only the spatial attention is considered in most of the works to extract the distinctiveness features. To learn the discriminative information from the images, we propose a channel-wise attention to learn the representations for person re-ID. In the rest part, we give a detailed description of the proposed method in Section 3.

## 3. Our Method

In this part, an overview of the proposed architecture is introduced, then we describe the proposed method for person re-ID.

### 3.1. Framework Overview

The framework for person re-ID is illustrated in Fig. 2. To obtain a robust feature representation, a novel network PGAN, for person re-ID, is designed. In the PGAN framework, we design a deep hybrid convolutional neural network and transformer structure to extract the distinctive features, which consists of a channel-wise attention mechanism, a residual block, and cswin transformer. In the following, a brief introduction of the CNN method is introduced.

### 3.2. CNN

Recently, the CNN has obtained great attentions in the DL community, which was originally proposed by Fukushima [8] based on “Neocognitron”, and was motivated from Hubel and Wiesel’s hierarchical receptive field of the visual cortex. Subsequently, LeCun

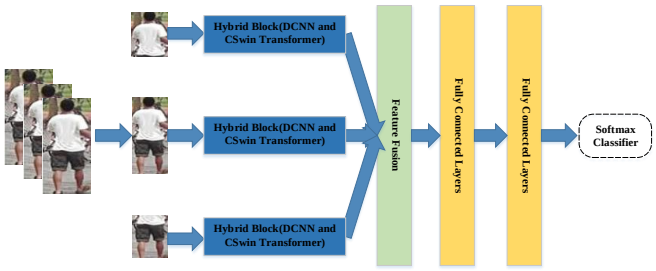


**Fig. 2.** The proposed deep architecture PGAN, in which Hybrid Block (DCNN and CSwin Transformer) is adopted to extract the deep features from the images and patches.

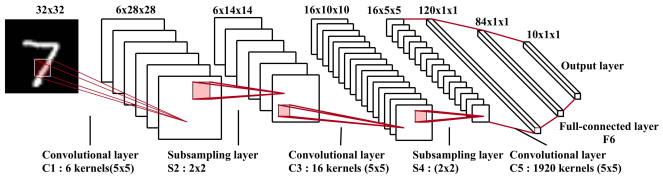


**Fig. 3.** LeNet-5 network. From [17]. The network contains three types of layers, i.e., convolutional, pooling, and fully-connected layers.

et al. [17] developed a framework for document recognition. In our work, we use LeNet-5 as an example to illustrate the architecture. Three layers is consisted in the LeNet-5: i.e., convolutional, pooling, and fully-connected layers. As illustrated in Fig. 3, we can discover that the convolution layer consists of a series of convolution kernels to compute the feature maps. The pooling layers are adopted to depress the spatial resolution of the feature maps. Commonly, the fully connect layers are used to generate the high-level feature representations [31]. Specifically, all the neurons of the fully-connected layers are used in the previous layer to link each single neuron of the current layer and obtain the output information. In some situations, the fully-connected layers can be modified with a $1 \times 1$ convolution layer to obtain the discriminative feature representations [22]. The main point of CNNs is that all the receptive fields share weights in a layer, needing only several parameters compared to fully-connected neural networks. Up to now, numerous CNN architectures have been designed to mitigate studies on image classification studies, i.e., AlexNet [16], ResNet [10], and VGGNet [31], etc.

### 3.3. Channel-wise Residual Block Based on CNN

The residual block structure of the channel-wise layer attention mechanism is illustrated in Fig. 4, which consists of two stacked basic residual network structures and a channel layer attention mechanism in the output of the last residual network structure. Suppose that the input feature map size is $\frac{C}{2} \times H \times W$ (where $C$ represents the number of feature map channels, and $H$ and $W$ represent the height and width of the feature map, respectively). Batch normalization is performed in the two convolutional layers. The number of feature map channels is doubled in the first convolutional layer of the first residual

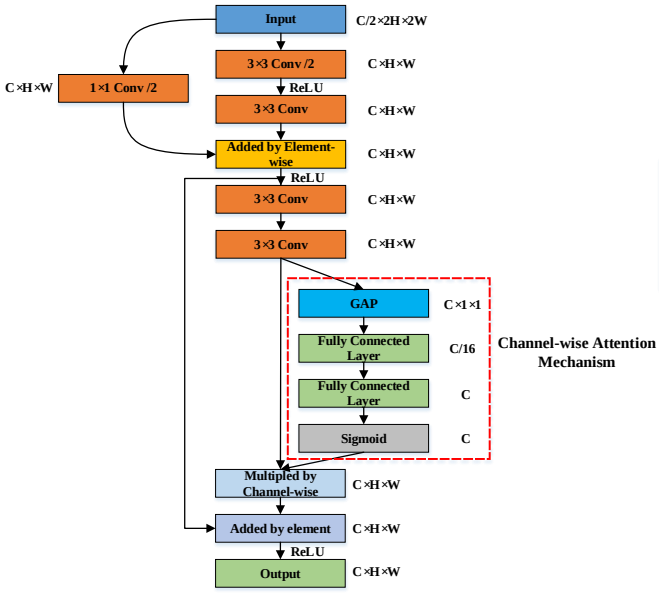


**Fig. 4.** The illustration of residual block.

network structure, and the feature map size is reduced. To ensure the obtaining of feature maps of the same size, we first use the $1 \times 1$ convolutional operation to double the number of the channels with a half number of the feature maps.

Firstly, the output feature map of the last convolutional layer is compressed at the spatial dimension, and the global average pooling is used to change the size of the feature map from the original $C \times H \times W$ to $1 \times 1 \times C$,

$$F_c(X) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X(i,j) \tag{1}$$

Where $X$ denotes the output feature map of the last convolutional layer.

Then, we learn the relationships among different channels, and the dimension of the channel features is reduced to 16. The ReLU is activated and then the size of the feature channels rises to the original feature channel dimension through a fully-connected layer. Advantages of this procedure include a higher non-linearity, a better fitting of the complex correlations among many channels, and reduction of the number of parameters as well as computations.

For the FC layer $F_c$, we have the following equation:

$$F_o = \mathrm{Re}LU(F_C \times w + b) \tag{2}$$

Where $F_o$ represents the output of the FC layers, $w$ denotes the weights of the FC layer, $b$ represents the bias of the FC layer.

For the output of the last fully-connected layer, sigmoid activation function is used for nonlinear transform to obtain the normalized weight of each channel within 0-1, and then

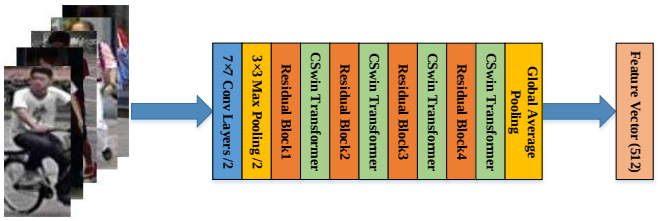


**Fig. 5.** The main architecture of the combination of Residual Block and Swin Transformer for feature learning.

**Table 1.** The size of feature maps.

| Network | The size of input feature map | The size of output feature map |
|---|---|---|
| Conv 7 × 7, stride 2 | 3×256×256 | 64×128×128 |
| Max Pooling 3×3, stride 2 | 64×128×128 | 64×64×64 |
| Residucal Block 1 + CSwin Transformer | 64×64×64 | 64×64×64 |
| Residucal Block 2 + CSwin Transformer | 64×64×64 | 128×32×32 |
| Residucal Block 3 + CSwin Transformer | 128×32×32 | 256×16×16 |
| Residucal Block 4 + CSwin Transformer | 256×16×16 | 512×8×8 |
| Global Average Pooling | 512×8×8 | 512×1×1 |

multiply back to the original channel features to complete the recalibration of the original features at the channel dimension. The procedure is written as:

$$\bar{X} = X \times Sigmoid(F_o) \tag{3}$$

Where $\bar{X}$ represents the feature map after recalibration, $X$ denotes the output feature map of the last convolutional layer, and $F_o$ is the output of the last FC layer.

The architecture of the deep convolutional neural network is shown in Fig. 5, which consists of a convolutional layer with kernel size of $7 \times 7$, a maximum pooling layer $3 \times 3$, residual blocks of four channel layer attention mechanisms and a global average pooling layer, with a final output feature vector of dimensions of 512. To reduce computational quantities and obtain large receptive fields near the input, the feature map size is downsampled twice consecutively using $7 \times 7$ convolution and $3 \times 3$ maximum pooling layers. Meanwhile, to ensure that not too much information is lost, the feature map size remains unchanged at the first residual block, and in the remaining three residual blocks, the feature map channels are doubled and the height as well as width is halved at the first convolution. For example, with the input feature map dimension of $3 \times 256 \times 256$, the input and output feature map dimension of each layer of the neural network is shown in Table 1.

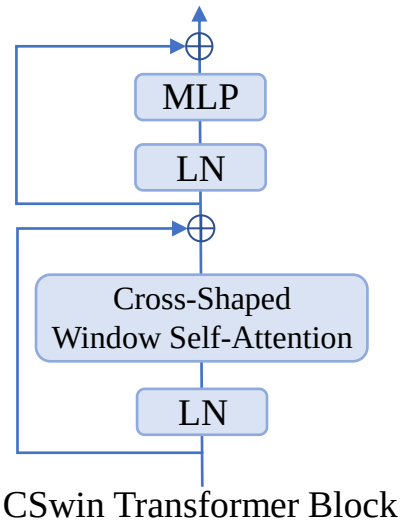


**Fig. 6.** The main architecture of CSwin Transformer.

### 3.4. PGAN

The ensemble pipeline of the proposed architecture is illustrated in Fig. 2. Firstly, we crop the images into the upper part of the body and the lower part of the body. Then the whole body and the upper as well as lower part of the body are input into the DCNN architecture to obtain the discriminative feature maps. To make a clear illustration, the three features are denoted as $F_w$, $F_u$ and $F_l$, respectively. After input the DCNN, the size of feature maps is $512$.

To learn the complimentary information in the three branches, we concatenate the $F_w$, $F_u$ and $F_l$ to generate the feature vector $F_{all}$,

$$F_{all} = [F_w, F_u, F_l] \tag{4}$$

Finally, we feed the feature vector into the FC layers, and perform the final classification performance.

### 3.5. Swin Transformer

Swin Transformer is proposed by [24] to model a hierarchical representation by starting from small-sized patches and gradually merging neighboring patches in deeper Transformer layers. As shown in Fig. 6, let assume an input of the size of $h \times w \times c$, Swin Transformer first adapt the input with the size of $\frac{hw}{M^2} \times M^2 \times c$ with window partitioning, where $\frac{hw}{M^2}$ denotes the number of windows (with the size $M \times M$). Then Swin Transformer performs self-attention operation with $h$ times in parallel on each window. Let assume a local window feature $Fea \in \mathbb{R}^{M^2 \times c}$, the $Q$, $K$ and $V \in \mathbb{R}^{M^2 \times d}$ are calculated by:

$$Q = FeaW_Q, K = FeaW_K, V = FeaW_V \tag{5}$$

Where $d$ is the query/key dimension, $W_Q, W_K, W_V$ represent the $query, keyandvalue$ matrices. The attention matrix $Attention(Q, K, V)$ is calculated by self-attention mechanism of the local window:

$$Attention(Q, K, V) = SoftMax(QK^T/\sqrt{d} + B)V \tag{6}$$

Where $B$ represents the learnable relative positional encoding. After that, a two-layer MLP with GELU activation followed. Then layer norm (LN) layer is added before multi-head self-attention (MSA) and MLP. The detailed description of Swin Transformer, please refer to [24].

Based on Swin Transformer, Dong et al. [6] propose a new transformer named CSwin Transformer as backbone for general-purpose vision tasks. Given an input image with the size of $H \times W \times 3$, we adopt the the overlapped convolutional token embedding ($7 \times 7$ convolution operatrion with stride 4) to generate $\frac{H}{4} \times \frac{W}{4}$ patch tokens with the dimension $C$. Based on the self-attention mechanism and positional embedding mechanism, CSwin Transformer block can be written as:

$$\begin{aligned} \hat{Z}^i &= CSWin - Attention(LN(Z^{i-1})) + Z^{i-1}, \\ Z^i &= MLP(LN(\hat{Z}^i)) + \hat{Z}^i \end{aligned} \tag{7}$$

where $Z^i$ represents the output of the $i-$ th CSWin Transformer block.

## 4. Experiments

### 4.1. Datasets

To make a fair comparison with the existing works, we adopted three databases to valid our proposed method. To verify the accuracy of the introduced architecture, extensive experiments were performed on the three benchmark database, i.e., Market-1501 [50], DukeMTMC-ReID [29,54] and CUHK03-NP [19,55]. Market-1501 consists of 12,936 images from 751 different identities. Query and gallery set consisted of 3,368 and 19,732 images from another 750 identities. DukeMTMC includes 16,522 data samples from 702 identities, including 2,228 and 17,661 images for the query and gallery set, respectively. CUHK03-NP is a subset of CUHK03, which included two series of data, i.e., labeled and detected images. The detected set of CUHK03 includes 7,365, 1,400 and 5,332 images for the training, query, and gallery partition, respectively. The labeled set of CUHK03 includes for the training, query and gallery partition of the number of 7,368, 1,400 and 5,328, respectively.

### 4.2. Experimental Settings and Evaluation Standards

### 4.3. Implementation Details:

To obtain a fast convergence for the deep models, an Adam optimizer with the parameters $\beta_1$, $\beta_2$ and $\varepsilon$ value 0.9, 0.999 and $1 \times 10^{-8}$ was used. The parameters of Adam had a momentum of 0.9, and a number of epochs of 50. The learning rate of 0.001 is set on the CUHKO-NP and Market-1501 dataset. While a learning rate of 0.0001 is set on the DukeMTMC-ReID dataset. The batch size was set as 32 on 1080Ti GPU. Pytorch platform is used in our experiment. To address the issues of overfitting, an early stop strategy is also adopted. The training process took 24 hours.

### 4.4. Results

In this part, we describe the performance for person re-ID. The performance of PGAN is introduced, and then discuss the results of person re-ID with those state-of-the-art methods.

**Table 2.** Performance of PGAN on the three benchmark databases.

| Methods | CUHK0-NP | | | | DukeMTMC-ReID | | Market-1501 | |
|---|---|---|---|---|---|---|---|---|
| | Labeled | | Detected | | | | | |
| | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP |
| | 83.1 | 81.8 | 79.4 | 80.3 | 92.5 | 83.4 | 97.0 | 91.3 |

**Performance of PGAN** The performance of PGAN on Market-1501, DukeMTMC-ReID and CUHK03-NP are illustrated in Table 2. From Table 2, one can see that comparable performances are obtained on three databases. Based on the CUHK0-NP dataset, we obtain the a mAP of 80.2% and 79.1% of the labeled and detected dataset, respectively. For the performance of DukeMTMC-ReID dataset, we obtain a mAP of 82.9% for person re-ID. For the Market-1501 dataset, we obtain the mAP of 90.2% for person re-ID. In our study, we try to use a patch-network and a global-network to extract the discriminative features of person re-ID. The effectiveness of the proposed method is also demonstrated from the videos.

**Table 3.** Comparison with the state of the art methods on CUHK03-NP.

| CUHK03-NP (%) | | | | | |
|---|---|---|---|---|---|
| Methods | Ref | Labeled | | Detected | |
| | | R-1 | mAP | R-1 | mAP |
| BoW+XQDA [50] | ICCV15 | 7.9 | 7.3 | 6.4 | 6.4 |
| SVDNet [33] | ICCV17 | – | – | 41.5 | 37.3 |
| DaRe(De)+RE [37] | CVPR18 | 66.1 | 61.6 | 63.3 | 59.0 |
| MLFN [3] | CVPR18 | 54.7 | 49.2 | 52.8 | 47.8 |
| HA-CNN [20] | CVPR18 | 44.4 | 41.0 | 41.7 | 38.6 |
| PCB+RPP [34] | ECCV18 | – | – | 63.7 | 57.5 |
| Mancs [33] | ECCV18 | 69.0 | 63.9 | 65.5 | 60.5 |
| CASN+PCB [52] | CVPR19 | 73.7 | 68.0 | 71.5 | 64.4 |
| RGA [48] | CVPR20 | 81.1 | 77.4 | 79.6 | 74.5 |
| NFormer [36] | CVPR22 | 80.6 | 79.1 | 79.0 | 76.4 |
| Ours | | 83.1 | 81.8 | 79.4 | 80.3 |

**Comparison with the State-of-the-Art Methods** Table 3, 4 and 5 illustrate the comparable accuracy with the state-of-the-art methods for person re-ID. Based on these observations, one can find that comparable results are obtained through our proposed method,

**Table 4.** Comparison with the state of the art methods on DukeMTMC.

| DukeMTMC-ReID (%) | | | |
|---|---|---|---|
| Methods | Ref | R-1 | mAP |
| BoW+kissme [50] | ICCV15 | 25.1 | 12.2 |
| SVDNet [33] | ICCV17 | 76.7 | 56.8 |
| DaRe(De)+RE [37] | CVPR18 | 80.2 | 64.5 |
| MLFN [3] | CVPR18 | 81.0 | 62.8 |
| KPM [30] | CVPR18 | 80.3 | 63.2 |
| HA-CNN [20] | CVPR18 | 80.5 | 63.8 |
| DNN-CRF [4] | CVPR18 | 84.9 | 69.3 |
| PABR [32] | ECCV18 | 84.4 | 69.3 |
| PCB+RPP [34] | ECCV18 | 83.3 | 69.2 |
| Mancs [33] | ECCV18 | 84.9 | 71.8 |
| CASN+PCB [52] | CVPR19 | 87.7 | 73.7 |
| PSTA [38] | ICCV21 | 98.3 | 97.4 |
| NFormer [36] | CVPR22 | 90.6 | 85.7 |
| —- [25] | 2023 | 88.3 | 75.1 |
| Ours | | 92.5 | 83.4 |

**Table 5.** Comparison with the state of the art methods on Market-1501.

| Market-1501 (%) | | | |
|---|---|---|---|
| Methods | Ref | R-1 | mAP |
| BoW+kissme [50] | ICCV15 | 44.4 | 20.8 |
| SVDNet [33] | ICCV17 | 82.3 | 62.1 |
| DaRe(De)+RE [37] | CVPR18 | 89.0 | 76.0 |
| MLFN [3] | CVPR18 | 90.0 | 74.3 |
| KPM [30] | CVPR18 | 90.1 | 75.3 |
| HA-CNN [20] | CVPR18 | 91.2 | 75.7 |
| DNN-CRF [4] | CVPR18 | 93.5 | 81.6 |
| PABR [32] | ECCV18 | 91.7 | 79.6 |
| PCB+RPP [34] | ECCV18 | 93.8 | 81.6 |
| Mancs [33] | ECCV18 | 93.1 | 82.3 |
| CASN+PCB [52] | CVPR19 | 94.4 | 82.8 |
| RGA [48] | CVPR20 | 96.1 | 88.4 |
| NFormer [36] | CVPR22 | 95.7 | 93.0 |
| PPLR [5] | CVPR22 | 94.3 | 84.4 |
| —- [25] | 2023 | 94.8 | 84.8 |
| Ours | | 97.0 | 91.3 |

showing the capability of our architecture. The reason is that only the deep models are trained from scratch in our research through our proposed method. However, we leverage the attention mechanism to learn the informative patterns from the videos. Using a channel-wise attention mechanism with a patch-network and a global-network, the main features of person re-ID can be learned.

## 5. Conclusion

In the presented paper, a new framework is introduced, named PGAN, which can disentangle representations can be learned for person re-ID. Of the proposed architecture, a novel attention method named channel-wise attention mechanism is used to learn the informative representations through the patch-network and global network. In addition, CSwin Transformer is also adopted to learn the informative information from the residual blocks. Extensive experiments are performed on the three public datasets, i.e., Market-1501, DukeMTMC-reID and CUHK03-NP, demonstrating the efficiency of the proposed method. Next, we will explore the deep learning methods to extract the deep-learned features and robust models for the person re-ID task.

**Acknowledgments.** This work was supported by the Science and Technology Project of the State Grid Corporation of China (5700-202019186A-0-0-00), the Shaanxi Provincial Natural Science Foundation (grant 2022JQ-718), and the National Statistical Research Project (grant 2021LY037).

## References

1. Bai, Y., Jiao, J., Ce, W., Liu, J., Lou, Y., Feng, X., Duan, L.Y.: Person30k: A dual-meta generalization network for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2123–2132 (2021)
2. Bai, Z., Wang, Z., Wang, J., Hu, D., Ding, E.: Unsupervised multi-source domain adaptation for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12914–12923 (2021)
3. Chang, X., Hospedales, T.M., Xiang, T.: Multi-level factorisation net for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2109–2118 (2018)
4. Chen, D., Xu, D., Li, H., Sebe, N., Wang, X.: Group consistent similarity learning via deep crf for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8649–8658 (2018)
5. Cho, Y., Kim, W.J., Hong, S., Yoon, S.E.: Part-based pseudo label refinement for unsupervised person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7308–7318 (2022)
6. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12124–12134 (2022)
7. Eom, C., Lee, W., Lee, G., Ham, B.: Is-gan: Learning disentangled representation for robust person re-identification. IEEE transactions on pattern analysis and machine intelligence (2021)
8. Fukushima, K., Miyake, S.: Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: Competition and cooperation in neural nets, pp. 267–285. Springer (1982)
9. Geng, Y., Hu, H.M., Zeng, G., Zheng, J.: A person re-identification algorithm by exploiting region-based feature salience. Journal of Visual Communication and Image Representation 29, 89–102 (2015)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Hou, R., Chang, H., Ma, B., Huang, R., Shan, S.: Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2014–2023 (2021)

12. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Feature completion for occluded person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
13. Hu, H.M., Fang, W., Zeng, G., Hu, Z., Li, B.: A person re-identification algorithm based on pyramid color topology feature. Multimedia Tools and Applications 76(24), 26633–26646 (2017)
14. Isobe, T., Li, D., Tian, L., Chen, W., Shan, Y., Wang, S.: Towards discriminative representation learning for unsupervised person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8526–8536 (2021)
15. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1062–1071 (2018)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 1097–1105 (2012)
17. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
18. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 369–378 (2018)
19. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 152–159 (2014)
20. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2285–2294 (2018)
21. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2197–2206 (2015)
22. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
23. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: Proceedings of the IEEE international conference on computer vision. pp. 350–359 (2017)
24. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
25. Ma, X., Lv, W., Zhao, M.: A double stream person re-identification method based on attention mechanism and multi-scale feature fusion. IEEE Access 11, 14612–14620 (2023)
26. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical gaussian descriptor for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1363–1372 (2016)
27. Meng, J., Zheng, W.S., Lai, J.H., Wang, L.: Deep graph metric learning for weakly supervised person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
28. Rao, H., Wang, S., Hu, X., Tan, M., Guo, Y., Cheng, J., Liu, X., Hu, B.: A self-supervised gait encoding approach with locality-awareness for 3d skeleton based person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
29. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. pp. 17–35. Springer (2016)
30. Shen, Y., Xiao, T., Li, H., Yi, S., Wang, X.: End-to-end deep kronecker-product matching for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6886–6895 (2018)

31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Computer Science (2014)
32. Suh, Y., Wang, J., Tang, S., Mei, T., Lee, K.M.: Part-aligned bilinear representations for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 402–419 (2018)
33. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3800–3808 (2017)
34. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European conference on computer vision (ECCV). pp. 480–496 (2018)
35. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. In: European conference on computer vision. pp. 791–808. Springer (2016)
36. Wang, H., Shen, J., Liu, Y., Gao, Y., Gavves, E.: Nformer: Robust person re-identification with neighbor transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7297–7307 (2022)
37. Wang, Y., Wang, L., You, Y., Zou, X., Chen, V., Li, S., Huang, G., Hariharan, B., Weinberger, K.Q.: Resource aware person re-identification across multiple resolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8042–8051 (2018)
38. Wang, Y., Zhang, P., Gao, S., Geng, X., Lu, H., Wang, D.: Pyramid spatial-temporal aggregation for video-based person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12026–12035 (October 2021)
39. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 79–88 (2018)
40. Wu, J., Yang, Y., Lei, Z., Wang, J., Li, S.Z., Tiwari, P., Pandey, H.M.: An end-to-end exemplar association for unsupervised person re-identification. Neural Networks 129, 43–54 (2020)
41. Xu, W., Liu, H., Shi, W., Miao, Z., Lu, Z., Chen, F.: Adversarial feature disentanglement for long-term person re-identification. In: Proceedings of the International Joint Conference on Artificial Intelligence (2021)
42. Yan, C., Pang, G., Bai, X., Liu, C., Xin, N., Gu, L., Zhou, J.: Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss. IEEE Transactions on Multimedia (2021)
43. Yan, Y., Qin, J., Ni, B., Chen, J., Liu, L., Zhu, F., Zheng, W.S., Yang, X., Shao, L.: Learning multi-attention context graph for group-based re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
44. Yang, Y., Tan, Z., Tiwari, P., Pandey, H.M., Wan, J., Lei, Z., Guo, G., Li, S.Z.: Cascaded split-and-aggregate learning with feature recombination for pedestrian attribute recognition. International Journal of Computer Vision pp. 1–14 (2021)
45. Yang, Y., Tiwari, P., Pandey, H.M., Lei, Z., et al.: Pixel and feature transfer fusion for unsupervised cross-dataset person reidentification. IEEE Transactions on Neural Networks and Learning Systems (2021)
46. Yang, Y., Zhang, T., Cheng, J., Hou, Z., Tiwari, P., Pandey, H.M., et al.: Cross-modality paired-images generation and augmentation for rgb-infrared person re-identification. Neural Networks 128, 294–304 (2020)
47. Yuan, L., Tian, Z.: Person re-identification based on color and texture feature fusion. In: International conference on intelligent computing. pp. 341–352. Springer (2016)
48. Zhang, Z., Lan, C., Zeng, W., Jin, X., Chen, Z.: Relation-aware global attention for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

49. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1077–1085 (2017)
50. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. pp. 1116–1124 (2015)
51. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
52. Zheng, M., Karanam, S., Wu, Z., Radke, R.J.: Re-identification with consistent attentive siamese networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5735–5744 (2019)
53. Zheng, W.S., Gong, S., Xiang, T.: Reidentification by relative distance comparison. IEEE transactions on pattern analysis and machine intelligence 35(3), 653–668 (2012)
54. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3754–3762 (2017)
55. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1318–1327 (2017)
56. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Learning generalisable omni-scale representations for person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)

**Liu Jing**, Female, PhD in Engineering, professor, works in Weinan Normal University. Her research interests include Computational photography and Light field imaging and processing.

**Zhou Guoqing**, Male, PhD in Engineering, associate professor, master's supervisor, and visiting scholar at Johns Hopkins University (JHU). His research directions include theories and issues in international frontier fields such as machine vision, computational photography, and global optimization theory.

*Received: August 29, 2023; Accepted: February 07, 2024.*