AE-Net: An efficient appearance-enriched neural network with foreground enhancement for person re-identification

Shangdong Zhu[a], Yunzhou Zhang[b], Yixiu Liu[b], Yu Feng[b], Sonya Coleman[c], Dermot Kerr[c]

[a]*Faculty of Robot Science and Engineering, Northeastern University, Shenyang 110819, China*
[b]*College of Information Science and Engineering, Northeastern University, Shenyang 110819, China*
[c]*Intelligent Systems Research Centre, Ulster University, Derry BT48 7JL, UK*

Abstract

Person re-identification (Re-ID) in environments subject to intensive appearance and background variations due to seasons, weather conditions, illumination and human factors is a challenging task. A wide variety of existing algorithms address this problem either for appearance changes or background clutter, but neglect to explore a powerful framework to consider solving both cases simultaneously. To overcome this limitation, this research introduces an efficient appearance-enriched neural network (AE-Net) with foreground enhancement based on generative adversarial nets (GANs) and an attention mechanism to enrich the appearance of person images while suppressing the influence of the background. Specifically, a channel-grouped convolution and squeeze weighted (CGCSW) module is first proposed to extract the powerful feature representation of individuals. Secondly, a foreground-enhanced and background-suppressed (FEBS) module is proposed to enhance the foreground of individual samples while weakening the impact of the background. Thirdly, A stage-wise consistency loss is presented to enable our model maintain consistent foreground-enhanced and background-suppressed stages. Finally, this study evaluates the proposed method and compares it with state-of-the-art approaches on three public datasets. The experimental results demonstrate the effectiveness and improvements achieved by using the presented architecture.

Keywords: Person re-identification, Appearance-enriched neural network, Channel-grouped convolution and squeeze weighted module, Foreground-enhanced and background-suppressed module.

1. Introduction

Person re-identification (Re-ID) [1] aims to match a person across non-overlapping camera views, and has recently attracted widespread attention [2, 3, 4, 5, 6]. It is an exceedingly important domain of computer vision, due to its potential extensive application prospects in video surveillance [7, 8], autonomous driving [9], etc. Although great progress has been achieved in person Re-ID tasks, it still remains a significant challenge in real-world environments with drastic changes in the background and individuals appearance (as shown in Fig. 1). On the one hand, it would be quite difficult to identify a person whose visual appearance changes between camera views. On the other hand, because of the variation in surveillance cameras and environmental changes, the subtle differences of individuals usually causes difficulty in the identification task. Therefore, it is essential to design an effective model to learn accurate representations that are robust to variations in appearance and environment.

Traditional person Re-ID methods [10, 11] learn features directly from local regions of the person, and thus it is difficult to obtain features that are robust to significant changes such as weather variations, illumination, viewpoint changes and pose changes. Deep learning algorithms have made significant progress over local regions for re-identification tasks. There are three mainstream deep learning methods based on Convolutional Neural Networks (CNNs) to realize person Re-ID. Several algorithms [12, 13] based on CNNs learn body regions to identity features which are obtained by either part area detection, or key points and pose estimation. Some researchers [14, 15, 16, 17] designed various methods based on CNNs to extract globally deep features from the whole images. Many CNNs-based attention strategies have been reported for reinforcement representation learning, such as the channel-wise feature re-weighting [18, 19, 20] and pixel level attention [20]. Nevertheless, various existing CNN-based methods pay so much attention to the information extraction of human body parts in the image that they have ignored the influence of the background, whereas enhancing the body region information while suppressing the background may bring higher performance improvements. Although some studies such as [21, 22] reveal the effectiveness of extracting features from the foreground body region rather than the background area, the off-line mask acquisition method and the simple removal of background clutter deviate further from real world scenarios. In this research, we focus on end-to-end feature representation enrichment of the individual body in the image while suppress the corresponding complex background at the feature-level. We aim to accurately separate the feature information that focuses on the person body area and the background region so as to selectively perform enhancement and weakening operations.

With the rapid development and application of the Generative

Adversarial Nets (GANs) [23], their powerful ability to obtain a large amount of augmented data can be considered as a potential direction to enhance robustness against appearance variations thus making it an appropriate choice for person Re-ID. Several works [24, 25, 26] employ unconditional GANs to generate augmented person examples and assign reasonable label distributions to them to assist in improving the performance of person Re-ID. Some researchers explore individual pose conditioned GANs to provide adequate pose coverage to train a robust person Re-ID system. The work in [27] indicates the effectiveness of unified networks which jointly couple generative and discriminative learning, however lack of high degree of attention to the background and foreground still results in low qualities of generated samples.

In order to overcome this limitation, we propose an efficient appearance-enriched neural network with foreground enhancement based on an attention mechanism and GANs to enhance the foreground of person images while suppressing the influence of the background for person Re-ID tasks. Unlike the aforementioned approaches that only involve the body region or background clutter elimination, this paper aims to learn both foreground enhancement and background suppression at the feature-level without resorting to pseudo labels of the generated samples. Fig. 2 illustrates an overview of the proposed framework which contains two main modules. The first module is based on a channel-grouped convolution and squeeze weighted (CGCSW) approach (embedded in the Strong Feature Representation (SFR) network) that has the capability to learn powerful feature representation of the individuals. This research mainly attempts to explore rich features that contain person information based on the channel and spatial level attention mechanism. Secondly, a feature separation module is presented to expand the distance between the feature information related to the individual body region and the background area, so as to selectively perform enhancement and weakening operations.

To summarize, the main contributions of this study are:

- A channel-grouped convolution and squeeze weighted (CGCSW) module is introduced to explore rich features of the person appearance information.
- A foreground-enhanced and background-suppressed (FEBS) module is presented to enhance the foreground of person samples while weakening the influence of the background for the person Re-ID task.
- A consistency loss is utilized to enable our model to maintain consistency between the foreground-enhanced and background-suppressed stages.
- Extensive experiments are conducted on three representative datasets that demonstrate the effectiveness and superiority of the proposed approach over the state-of-the-art methods.

2. Related work

In this section, we review the existing works related to person re-identification (Re-ID) from three aspects: mask-guided models for person Re-ID, visual attention mechanism-based person Re-ID and GAN-based person Re-ID.

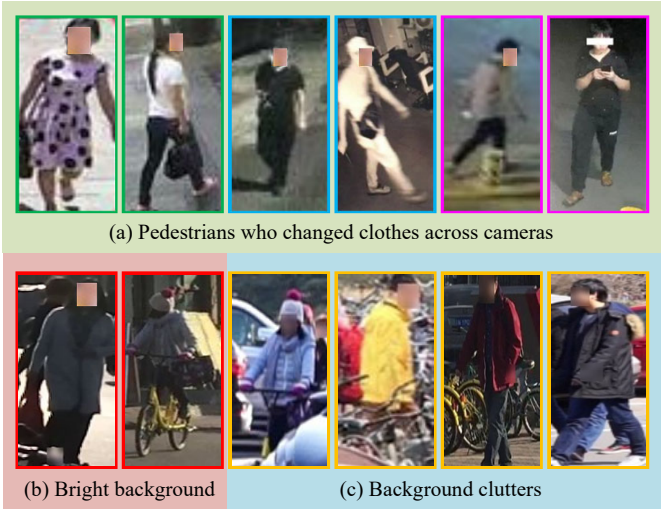


Figure 1: Examples when (a) persons changed clothes across cameras; (b) the background is extremely bright; (c) individuals with quite complex environments.

2.1. Mask-guided person Re-ID

As the topic of segmentation has attracted increasing attention recently, its application in person Re-ID has also been particularly widespread. A Mask-guided Contrastive Attention Model (MGCAM) is introduced by Song et al. [21] to extract robust and discriminative features which are invariant to background clutter. The authors are the first to apply the binary mask to the person Re-ID task successfully. Chen et al. [28] present the Mask-Guided Two-Stream CNN model (MGTS) to enhance the representation by merging one stream from the original sample and make full use of another separate stream from the foreground as the emphasis message. This further provides the model with the ability to extract more representative features of each persons identity. In [29], Cai et al. propose a multi-scale body-part mask-guided attention network (MMGA), which studies the attention of part-body regions and entire-body areas to assist in extracting local and global features. In addition, the approach does not require the use of a mask during inference, which makes it particularly effective. Chen et al. [30] report the Confidence Weighted Stream Attention (CWSA) algorithm to re-weight the two-stream model, which considers the relative importance of the two streams in a more rigorous manner.

The majority of these algorithms are primarily concerned with the correspondence between a single sample image and a label, while ignoring the rich global mutual information in an entire dataset. To solve this issue, Bao et al. [31] introduced the Masked Graph Attention Network (MGAT) to employ the rich global mutual information among extracted features. Unlike most previous algorithms that primarily employ a mask-guided approach to extract discriminative and robust features which are invariant to background clutter, this research adopts an appearance enrichment with foreground enhancement technique to im-

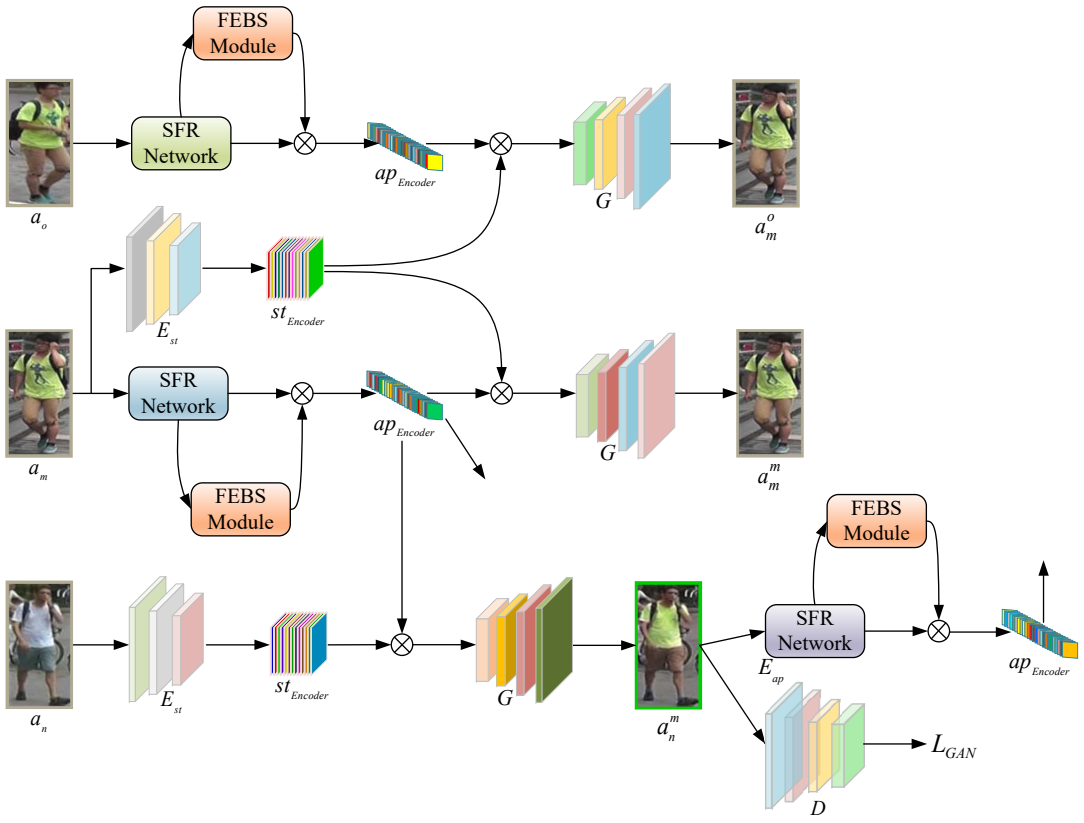


Figure 2: Overview of the proposed appearance-enriched framework with foreground enhancement (AE-Net), where our proposed channel-grouped convolution and squeeze weighted (CGCSW) module (embedded into the SFR network) and foreground-enhanced and background-suppressed (FEBS) module are embedded into the architecture.

prove the performance of the person Re-ID task without the use of any external mask.

2.2. Visual Attention Mechanism-based Person Re-ID

Motivated by the recently successful application of visual attention mechanisms for person Re-ID, a variety of works [32, 33, 34, 35, 36, 5] have been presented to achieve accurate individual identity matching through precise body region attention. Zheng et al. [32] design the Consistent Attentive Siamese Network (CASN), which is a novel siamese learning framework driven by attention for view-invariant representation learning to achieve robust cross-view matching. In [33], the Batch DropBlock (BDB) Network is reported by Dai et al., which is a two branch architecture containing both a feature dropping branch and a traditional global branch to weaken the suppressed attentive local features during training. Fu et al. [35] first propose the High-Order Attention (HOA) module to utilize high-order statistical information in the attention mechanism, and further capture subtle differences among individuals and provide discriminative attention proposals. Li et al. [36] present the harmonious attention network (HAN) architecture to learn both hard area attention and soft pixel attention simultaneously. They further research feature representations to maximize the complementary correlated information between feature discrimination and attention selection in a compact framework. In order to address the problem of recognition difficulty caused by subtle differences in appearance of different persons, Qian et al. [5] introduce a multi-scale deep learning network (MuDeep) for identity recognition. Although these works have achieved a considerable performance improvements, they still do not meet the current requirements of the person Re-ID task.

2.3. GAN-based Person Re-ID

In [23], Goodfellow et al. propose the Generative Adversarial Nets (GANs) which are able to generate similar examples based on a deep understanding of the network. With increasing progress and application of GANs, various research [24, 25, 26, 37, 38, 39, 40, 41] have made use of their powerful capability to obtain augmented data and enhance robustness against input variations. Zheng et al. [26] design a semi-supervised framework with label smoothing regularization for outliers (LSRO), which assigns the unmarked samples with a uniform label distribution and regularizes the neural network in the training process. Qian et al. [37] introduce a person sample generated architecture named pose-normalization GAN (PN-GAN) to generate pose-normalized samples with the advantages of identity preservation, realism and posture controllability. In order to more efficiently utilize the generated samples for an improved feature learning and person Re-ID performance, Huang et al. [24] and Ding et al. [25] report the Multi-pseudo Regularized Label (MpRL) and Feature Affinity-based Pseudo Labeling (FAPL) algorithms to assign more accu-

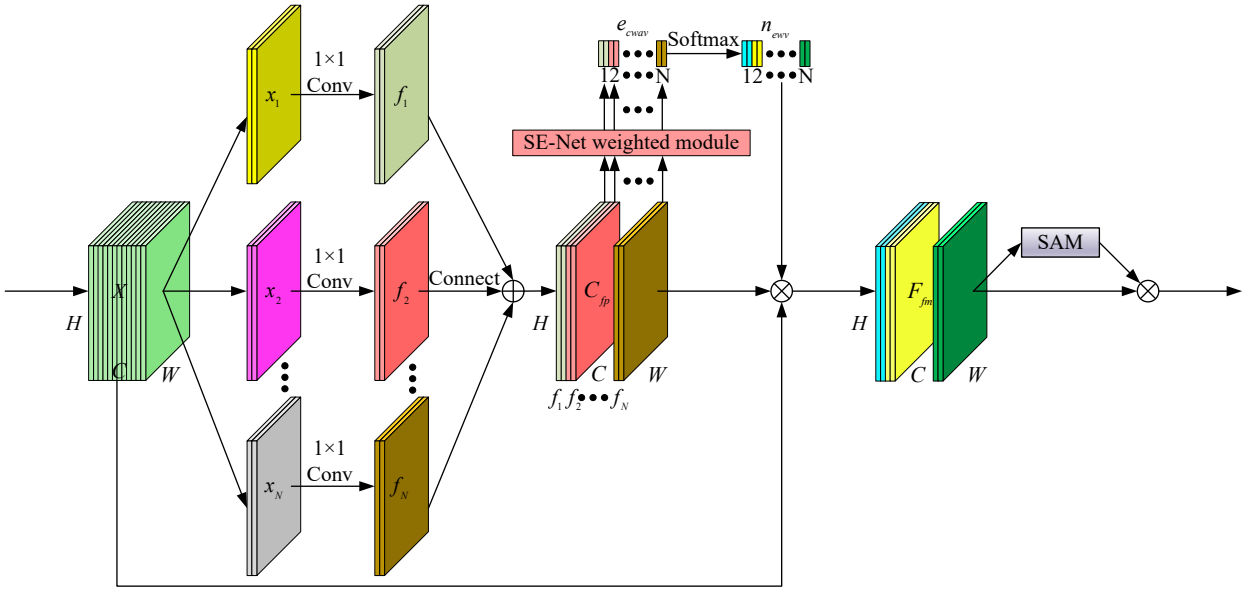


Figure 3: Channel-grouped convolution and squeeze weighted (CGCSW) module.

rate labels to the generated artificial samples. Zheng et al. [27] present an architecture that integrates generative and discriminative learning in a unified network named DG-Net. Although this paper adopts DG-Net as the baseline approach, our framework mainly focuses on enhancing the foreground of individual samples while weakening the influence of the background and simultaneously conducting strong feature representation operations, which is very different from the original DG-Net [27].

3. Technical approach

We introduce the appearance-enriched neural network with foreground enhancement (AE-Net) in detail for the person Re-ID task. Specifically, we first report the channel-grouped convolution and squeeze weighted (CGCSW) module to obtain powerful feature representations of the persons. Secondly, we introduce the foreground-enhanced and background-suppressed (FEBS) module to enhance the foreground of the person samples while weakening the influence of the background on the person Re-ID task. Finally, consistency loss is presented to optimize the FEBS module. The overall framework of the developed AE-Net is illustrated in Fig. 2.

3.1. Channel-grouped Convolution and Squeeze Weighted module

In order to extract powerful feature representations of individuals, this study designs a channel-grouped convolution and squeeze weighted (CGCSW) module (shown in Fig. 3). The proposed CGCSW module is partly inspired from the Efficient Pyramid Squeeze Attention (EPSANet) approach [42], which designs a Pyramid Squeeze Attention (PSA) module which builds a more efficient and effective channel attention mechanism. Nevertheless, unlike EPSANet, we do not adopt the Squeeze and Concat (SPC) module which obtains the multi-scale feature map through the channel-wise attention vector. Instead, we directly adopt the 1×1 convolution operation to reduce the number of parameters and add non-linear features while maintaining the dimensionality of the input. This makes full use of the input image to prevent the suppression of input features. Furthermore, we adopt the Spatial Attention Module (SAM) component of the Convolutional Block Attention Module (CBAM) [43] to focus on the spatially informative part of the image.

Fig. 3 illustrates the CGCSW module which consists primarily of five steps. Firstly, the channel-grouped convolution feature maps are obtained through the 1×1 convolution operation, and then all the channel-grouped convolution feature maps are connected in sequence. Secondly, in order to extract the channel attention of the connected feature map, the excited channel-wise attention vector is obtained by executing the SE-Net [44] weighted module. Thirdly, the Softmax function is applied to normalize the channel-wise attention vector to obtain the normalized excitation-weighted attention vector. Fourthly, the element-wise multiplication and channel-wise multiplication are applied to the input, the normalized excitation-weighted attention vector and the connected feature map to obtain the fusion feature map. Finally, the SAM is applied to the fusion feature map to obtain the output feature map, which is sensitive to the channel and spatial features.

Specifically, as illustrated in Fig. 3, suppose $\boldsymbol{X} \in \mathbb{R}^{C \times H \times W}$ represents the input feature map, where C, H and W represent the channel number, spatial height and width of the input feature map, respectively. The CGCSW module first divides the input feature map $\boldsymbol{X}$ into N groups according to the channel dimension, i.e., $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]$ where the channel-grouped feature map $\boldsymbol{x}_i \in \mathbb{R}^{C/N \times H \times W}$ (we set the value of C/N to 2, which means that each $\boldsymbol{x}_i$ has two channels) and $i = 1, 2, \cdots, N$. To maintain the dimensionality of the input image while reducing the number of parameters and adding non-linear features, this approach performs the 1×1 convolution procedure

on each $\boldsymbol{x}_i$ to obtain the channel-grouped convolution feature map $\boldsymbol{f}_i \in \mathbb{R}^{2\times H\times W}$ $(i = 1, 2, \cdots, N)$, and then connects all $\boldsymbol{f}_i$ to obtain the connected feature map $\boldsymbol{C}_{fp} \in \mathbb{R}^{C\times H\times W}$, which can be formulated as follows:

$$\begin{aligned}\boldsymbol{C}_{fp} &= \mathrm{CON}(\boldsymbol{f}_i) \\ &= [\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_N],\ i = 1, 2, \cdots, N,\end{aligned} \tag{1}$$

where CON represents the connection process to obtain the $\boldsymbol{C}_{fp}$. This paper employs the excited channel-wise attention vector $\boldsymbol{e}_{cwav}$ (expressed in Eq. (5)) obtained by the SE-Net weighted module to extract the channel attention of the $\boldsymbol{C}_{fp}$. Suppose $\boldsymbol{f}_i = [\boldsymbol{f}_{i1}, \boldsymbol{f}_{i2}]$, $\boldsymbol{f}_{i1}, \boldsymbol{f}_{i2} \in \mathbb{R}^{1\times H\times W} (i = 1, 2, \cdots, N)$, then the $\boldsymbol{C}_{fp}$ can be reformulated as follows:

$$\begin{aligned}\boldsymbol{C}_{fp} &= [\boldsymbol{f}_{11}, \boldsymbol{f}_{12}, \boldsymbol{f}_{21}, \boldsymbol{f}_{22}, \cdots, \boldsymbol{f}_{N1}, \boldsymbol{f}_{N2}] \\ &= [\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_{2N}],\end{aligned} \tag{2}$$

where each $\boldsymbol{f}_j \in \mathbb{R}^{1\times H\times W}$ $(j = 1, 2, \cdots, 2N)$. Thus, the channel vector $\boldsymbol{c}_v \in \mathbb{R}^{C\times 1\times 1}$ after compressing $\boldsymbol{C}_{fp}$ in the channel direction, can be expressed as follows:

$$\boldsymbol{c}_v = [c_1, c_2, \cdots, c_{2N}], \tag{3}$$

where each c_j $(j = 1, 2, \cdots, 2N)$ is the channel-shrunken element (obtained by the global average pooling to squeeze global information) which can be calculated as follows:

$$c_j = \frac{1}{HW}\sum_{m=1}^{H}\sum_{n=1}^{W}\boldsymbol{f}_j(m, n), \tag{4}$$

where (m, n) represents the element of the feature map $\boldsymbol{f}_j$. Then the excited channel-wise attention vector $\boldsymbol{e}_{cwav} \in \mathbb{R}^{C\times 1\times 1}$ can be expressed as follows:

$$\begin{aligned}\boldsymbol{e}_{cwav} &= [e_1, e_2, \cdots, e_{2N}] \\ &= \zeta(\boldsymbol{w}_2\xi(\boldsymbol{w}_1\boldsymbol{c}_v)),\end{aligned} \tag{5}$$

where each e_i $(i = 1, 2, \cdots, 2N)$ is the i-th element of the $\boldsymbol{e}_{cwav}$, ζ and ξ denote the Sigmoid function and Rectified Linear Unit (ReLU) activation function [45], respectively, $\boldsymbol{w}_1 \in \mathbb{R}^{\frac{C}{\rho}\times C}$ and $\boldsymbol{w}_2 \in \mathbb{R}^{C\times\frac{C}{\rho}}$ are the parameters of two fully-connected (FC) layers, respectively and ρ is the reduction ratio which reduces the number of channels by the dimensionality-reduction layer (the first FC layer) and thus reduces the computation. Note ρ is preferentially chosen as 8, and a detailed explanation for this choice is provided in the ablation analysis in Section 4.5.2. To further establish long-term channel attention dependence and realize the information interaction among channel attention, the Softmax function η is adopted to normalize the weight of the channel attention information for the excited channel-wise attention vector $\boldsymbol{e}_{cwav}$. The normalized excitation-weighted vector $\boldsymbol{n}_{ewv} \in \mathbb{R}^{C\times 1\times 1}$ can be formulated as follows:

$$\begin{aligned}\boldsymbol{n}_{ewv} &= [n_1, n_2, \cdots, n_{2N}] \\ &= \eta(\boldsymbol{e}_{cwav}) \\ &= \eta(\zeta(\boldsymbol{w}_2\xi(\boldsymbol{w}_1\boldsymbol{c}_v))),\end{aligned} \tag{6}$$

where each n_i $(i = 1, 2, \cdots, 2N)$ can be defined as follows:

$$\begin{aligned}n_i &= \eta(e_i) \\ &= \frac{\exp(e_i)}{\sum_{i=1}^{N}\exp(e_i)}.\end{aligned} \tag{7}$$

Thus, the final fusion feature map $\boldsymbol{F}_{fm} \in \mathbb{R}^{C\times H\times W}$ which fuses the features of the input feature map $\boldsymbol{X}$, the connected feature map $\boldsymbol{C}_{fp}$ and the normalized excitation-weighted vector $\boldsymbol{n}_{ewv}$ can be expressed as follows:

$$\boldsymbol{F}_{fm} = [n_1 \odot \boldsymbol{f}_1 \otimes \boldsymbol{x}_1, n_2 \odot \boldsymbol{f}_2 \otimes \boldsymbol{x}_2, \cdots, n_{2N} \odot \boldsymbol{f}_{2N} \otimes \boldsymbol{x}_{2N}], \tag{8}$$

where the $\odot$ and $\otimes$ represent the channel-wise multiplication and element-wise multiplication, respectively. The SAM is applied to the $\boldsymbol{F}_{fm}$ to focus on the spatially informative part. Thus, the output feature map $\boldsymbol{O}_{fm} \in \mathbb{R}^{C\times H\times W}$ which fuses both the channel and spatial information of the input feature map can be formulated as follows:

$$\boldsymbol{O}_{fm} = \zeta(\mathrm{Conv}_7([\mathrm{AP}(\boldsymbol{F}_{fm}), \mathrm{MP}(\boldsymbol{F}_{fm})])) \otimes \boldsymbol{F}_{fm}, \tag{9}$$

where Conv_7 denotes the convolution with the filter size of 7×7. The AP and MP represent the Average Pooling and the Max Pooling operations, respectively. On the basis of the proposed CGCSW module, which could obtain the powerful feature representation of persons, this paper will introduce the foreground-enhanced and background-suppressed learning in the next section.

3.2. Foreground-enhanced and background-suppressed learning

The complex background of person samples and the corresponding environmental changes could seriously affect the recognition accuracy of the person Re-ID task. Based a series of experiments that have been completed, we observe that the feature maps in the lower layers of the convolution neural networks (CNNs) represent the apparent properties, while the higher layers of the CNNs reflect the semantic attributes. Meanwhile, we observe that even in the final output feature map, different channels focus on background regions or foreground areas. Therefore, we determined that the Max Pooling operation (equivalent to enhancement) for the channels is concerned with foreground regions and the Average Pooling operation (equivalent to weakening) for the channels focuses on the background areas. This approach can actually obtain a significant foreground-enhanced and background-suppressed effect, and further weaken the affect of the background on any subsequent re-identification processes.

To reduce the influence of the background regions, this research introduces the foreground-enhanced and background-suppressed (FEBS) module (shown in Fig. 4) which can enhance the information in the foreground while weakening the background of person images. Fig. 4 shows the FEBS module consists of two branches which use the Stage 1 and Stage 4 outputs of the SFR network). These are assigned with similar tasks, i.e., both are employed for the foreground enhancement

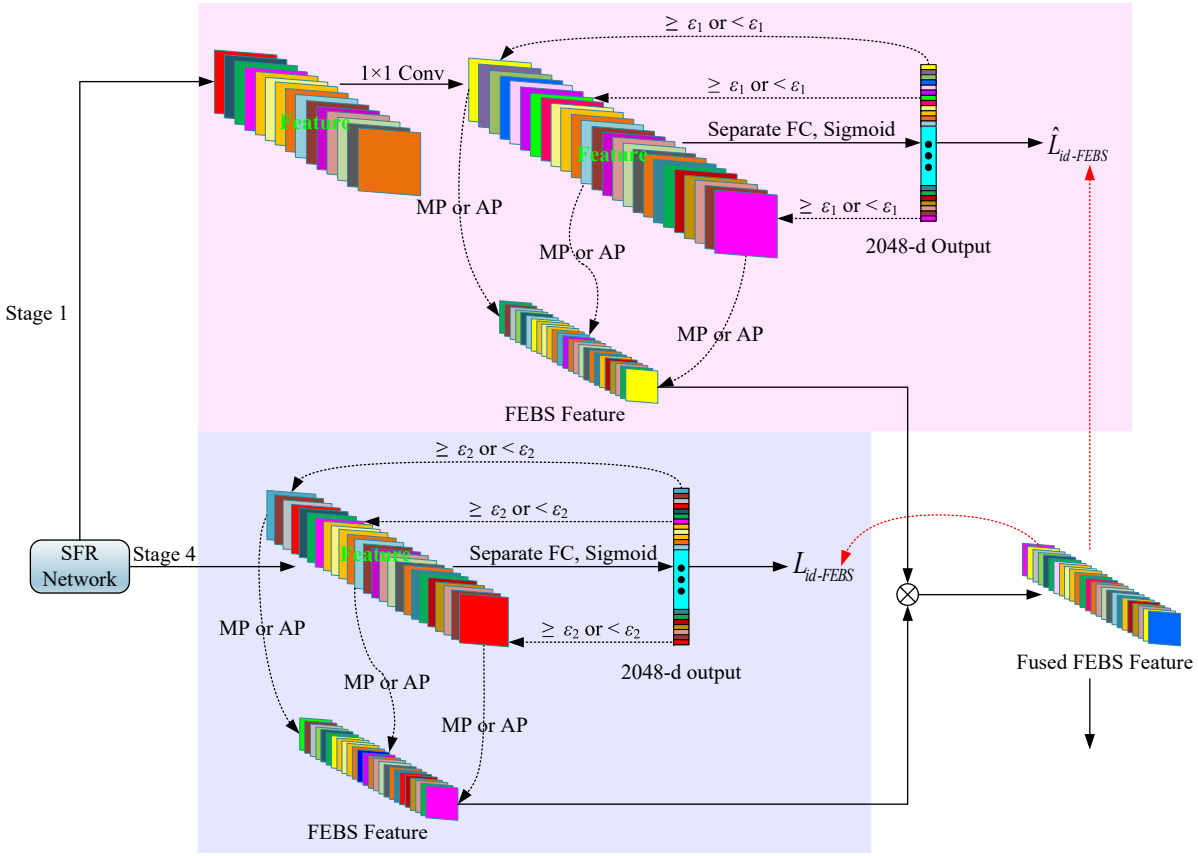


Figure 4: The diagram of the proposed foreground-enhanced and background-suppressed (FEBS) module.

and background suppression. Each of the two branches is connected in line with the fact that the lower layers contain apparent information and the higher layers contain semantic information. Combining these two types of information could further improve the recognition capability of the person Re-ID task. In general, the proposed FEBS module is primarily comprised of two steps. Firstly, the two branches are processed separately to obtain the FEBS features by using the separate FC layer and the Sigmoid activation function. Secondly, the FEBS features obtained from the two branches are fused as the final FEBS feature map.

More specifically, suppose $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \cdots, \tilde{\boldsymbol{x}}_{\tilde{C}}]$ and $\check{\boldsymbol{X}} = [\check{\boldsymbol{x}}_1, \check{\boldsymbol{x}}_2, \cdots, \check{\boldsymbol{x}}_{\check{C}}]$ are the outputs of Stage 1 and Stage 4 of the SFR Network, where $\tilde{\boldsymbol{X}} \in \mathbb{R}^{\tilde{C}\times\tilde{H}\times\tilde{W}}$, $\check{\boldsymbol{X}} \in \mathbb{R}^{\check{C}\times\check{H}\times\check{W}}$, $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{1\times\tilde{H}\times\tilde{W}}$ $(i = 1, 2, \cdots, \tilde{C})$ and $\check{\boldsymbol{x}}_i \in \mathbb{R}^{1\times\check{H}\times\check{W}}$ $(i = 1, 2, \cdots, \check{C})$ represent a single feature map. $\tilde{C}$ and $\check{C}$ are the number of channels of the $\tilde{\boldsymbol{X}}$ and $\check{\boldsymbol{X}}$. In order to obtain the FEBS features with the same dimensionality from both branches, the dimension of $\tilde{\boldsymbol{X}}$ in the channel direction needs to be increased by a 1×1 convolution. Hence, $\hat{\boldsymbol{X}}$ can be obtained after the channel direction is increased and is expressed as:

$$\begin{aligned}\hat{\boldsymbol{X}} &= \mathrm{Conv}_1(\tilde{\boldsymbol{X}}) \\ &= [\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \cdots, \hat{\boldsymbol{x}}_{\check{C}}],\end{aligned} \tag{10}$$

where $\hat{\boldsymbol{X}} \in \mathbb{R}^{\check{C}\times\tilde{H}\times\tilde{W}}$, Conv_1 denotes the convolution with the filter size of 1×1 and $\hat{\boldsymbol{x}}_i \in \mathbb{R}^{1\times\tilde{H}\times\tilde{W}}$ $(i = 1, 2, \cdots, \check{C})$ represents the i-th channel feature map of $\hat{\boldsymbol{X}}$. For both branches, the separate FC layer corresponding to each channel of $\hat{\boldsymbol{X}}$ and $\check{\boldsymbol{X}}$ is adopted, which is activated by the Sigmoid function. Thus, the 2048 dimensional outputs $\hat{\boldsymbol{s}} \in \mathbb{R}^{\check{C}\times1\times1}$ and $\check{\boldsymbol{s}} \in \mathbb{R}^{\check{C}\times1\times1}$ of the two branches can be formulated as follows:

$$\begin{aligned}\hat{\boldsymbol{s}} &= [\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_{\check{C}}] \\ &= \zeta(\hat{\boldsymbol{w}}_1(\mathrm{R}(\hat{\boldsymbol{x}}_1)), \hat{\boldsymbol{w}}_2(\mathrm{R}(\hat{\boldsymbol{x}}_2)), \cdots, \hat{\boldsymbol{w}}_{\check{C}}(\mathrm{R}(\hat{\boldsymbol{x}}_{\check{C}}))), \\ \check{\boldsymbol{s}} &= [\check{s}_1, \check{s}_2, \cdots, \check{s}_{\check{C}}] \\ &= \zeta(\check{\boldsymbol{w}}_1(\mathrm{R}(\check{\boldsymbol{x}}_1)), \check{\boldsymbol{w}}_2(\mathrm{R}(\check{\boldsymbol{x}}_2)), \cdots, \check{\boldsymbol{w}}_{\check{C}}(\mathrm{R}(\check{\boldsymbol{x}}_{\check{C}}))),\end{aligned} \tag{11}$$

where each of the $\hat{s}_i$ and $\check{s}_i$ $(i = 1, 2, \cdots, \check{C})$ represents the i-th elements of the $\hat{\boldsymbol{s}}$ and $\check{\boldsymbol{s}}$, $\mathrm{R}(\hat{\boldsymbol{x}}_i) \in \mathbb{R}^{\tilde{H}\tilde{W}\times1\times1}$ and $\mathrm{R}(\check{\boldsymbol{x}}_i) \in \mathbb{R}^{\check{H}\check{W}\times1\times1}$ denote channel flat vectors of $\hat{\boldsymbol{X}}$ and $\check{\boldsymbol{X}}$ respectively, $\hat{\boldsymbol{w}}_i \in \mathbb{R}^{1\times\tilde{H}\tilde{W}}$ and $\check{\boldsymbol{w}}_i \in \mathbb{R}^{1\times\check{H}\check{W}}$ $(i = 1, 2, \cdots, \check{C})$ represent the parameters of the separate FC layers and ζ denotes the Sigmoid activation function.

In order to make full use of the information fed back from the Sigmoid function and guide the two branches to obtain the FEBS features, we employ two hyperparameters ε_1 and ε_2, which represent the thresholds to determine whether to perform the Max Pooling (MP) operation for foreground enhancement or the Average Pooling (AP) operation for background suppression in the current channel. Therefore, the feature maps $\hat{o}_i$ and $\check{o}_i$ are obtained as follows:

$$\hat{\boldsymbol{f}}_i = \begin{cases} \text{MP} & \hat{s}_i \geq \varepsilon_1 \\ \text{AP} & \hat{s}_i < \varepsilon_1 \end{cases}, \ i = 1, 2, \cdots, \check{C}, \\ \check{\boldsymbol{f}}_i = \begin{cases} \text{MP} & \check{s}_i \geq \varepsilon_2 \\ \text{AP} & \check{s}_i < \varepsilon_2 \end{cases}, \ i = 1, 2, \cdots, \check{C}, \tag{12}$$

where the detailed verification process and explanation of the two hyperparameters ε_1 and ε_2 can be found in Section 4.5.3. Then, the two FEBS feature maps $\hat{\boldsymbol{F}} \in \mathbb{R}^{\check{C}\times H\times W}$ and $\check{\boldsymbol{F}} \in \mathbb{R}^{\check{C}\times H\times W}$, obtained by the two branches, can be formulated as follows:

$$\hat{\boldsymbol{F}} = [\hat{\boldsymbol{f}}_1, \hat{\boldsymbol{f}}_2, \cdots, \hat{\boldsymbol{f}}_{\check{C}}], \\ \check{\boldsymbol{F}} = [\check{\boldsymbol{f}}_1, \check{\boldsymbol{f}}_2, \cdots, \check{\boldsymbol{f}}_{\check{C}}], \tag{13}$$

where each of the $\hat{\boldsymbol{f}}_i \in \mathbb{R}^{1\times H\times W}$ and $\check{\boldsymbol{f}}_i \in \mathbb{R}^{1\times H\times W}$ ($i = 1, 2, \cdots, \check{C}$) denotes the i-th feature maps of the $\hat{\boldsymbol{F}}$ and $\check{\boldsymbol{F}}$ respectively.

This research fuses the two FEBS feature maps as the final fused FEBS feature map which can be expressed as:

$$\boldsymbol{F}_{fused} = \hat{\boldsymbol{F}} \odot \check{\boldsymbol{F}}, \tag{14}$$

where the $\odot$ denotes the channel-wise multiplication.

3.3. Stage-wise Consistency Optimization

To optimize the weights of the proposed FEBS module, this paper adopts identification loss to distinguish different identities. Thus, the self-identity losses for each of the two branches can be respectively formulated as:

$$\hat{\mathcal{L}}^s_{id\text{-}FEBS} = \mathbb{E}[-\log(p(b_m|\hat{a}_m))], \\ \check{\mathcal{L}}^s_{id\text{-}FEBS} = \mathbb{E}[-\log(p(b_m|\check{a}_m))], \tag{15}$$

where $\hat{a}_m$ and $\check{a}_m$ denote the input of the higher and lower branches, m is the number of samples, $b_m \in [1, L]$ indicates the identity label, L represents the number of identities and $p(b_m|\hat{a}_m)$ and $p(b_m|\check{a}_m)$ are predicted probabilities that $\hat{a}_m$ and $\check{a}_m$ belong to the ground-truth class b_m. In addition, the cross-identity losses of the two branches can be formulated as follows:

$$\hat{\mathcal{L}}^c_{id\text{-}FEBS} = \mathbb{E}[-\log(p(b_m|\hat{a}^n_m))], \\ \check{\mathcal{L}}^c_{id\text{-}FEBS} = \mathbb{E}[-\log(p(b_m|\check{a}^n_m))], \tag{16}$$

where $\hat{a}^n_m$ and $\check{a}^n_m$ now denote the input of the higher and lower branches corresponding to image n.

Since the $\hat{\mathcal{L}}^s_{id\text{-}FEBS}$ and $\check{\mathcal{L}}^s_{id\text{-}FEBS}$ are loss calculations for the same identity, we propose a stage-wise loss $\mathcal{L}^s_{id\text{-}FEBS}$ to combine the stage losses which can be formulated as follows:

$$\mathcal{L}^s_{id\text{-}FEBS} = \hat{\mathcal{L}}^s_{id\text{-}FEBS} + \check{\mathcal{L}}^s_{id\text{-}FEBS} + \|(\hat{\boldsymbol{P}}_s - \check{\boldsymbol{P}}_s)\|_2 \tag{17}$$

where $\hat{\boldsymbol{P}}_s$ and $\check{\boldsymbol{P}}_s$ are the vectors containing $p(b_m|\hat{a}_m)$ and $p(b_m|\check{a}_m)$, respectively. Similar to $\mathcal{L}^s_{id\text{-}FEBS}$, another stage-wise loss $\mathcal{L}^c_{id\text{-}FEBS}$ is determined as follows:

$$\mathcal{L}^c_{id\text{-}FEBS} = \hat{\mathcal{L}}^c_{id\text{-}FEBS} + \check{\mathcal{L}}^c_{id\text{-}FEBS} + \|(\hat{\boldsymbol{P}}_c - \check{\boldsymbol{P}}_c)\|_2 \tag{18}$$

where $\hat{\boldsymbol{P}}_c$ and $\check{\boldsymbol{P}}_c$ represent the vectors containing $p(b_m|\hat{a}^m_n)$ and $p(b_m|\check{a}^m_n)$, respectively.

In addition to the losses $\mathcal{L}^s_{id\text{-}FEBS}$ and $\mathcal{L}^c_{id\text{-}FEBS}$, this study follows the approach of the [27], and includes the loss function $\mathcal{L}_{DG\text{-}Net}$. Therefore, the total loss can be expressed as follows:

$$\mathcal{L}_{AE\text{-}Net} = \mathcal{L}^s_{id\text{-}FEBS} + \mathcal{L}^c_{id\text{-}FEBS} + \mathcal{L}_{DG\text{-}Net}. \tag{19}$$

4. Experimental evaluation

This study conducts experiments on three widely recognized large-scale person Re-ID datasets, including Market-1501 [46], DukeMTMC-reID [26] and MSMT17 [47], to validate the effectiveness of the proposed model. Firstly, we introduce these datasets and the corresponding evaluation metrics. Secondly, the experimental set-up details are presented. Thirdly, this paper compares the proposed model with state-of-the-art GAN-based person Re-ID methods and other Re-ID approaches. Fourthly, ablation experiments are used to verify the effectiveness of each component. Finally, we report the results from both quantitative and qualitative perspectives.

Table 1: Comparison of experimental results with the published state-of-the-art GANs-based methods on the Market-1501 and DukeMTMC-reID datasets. Rank-1, Rank-5, Rank-10 and mAP are listed. The best results among these methods are highlighted in bold.

Methods	References	Market-1501				DukeMTMC-reID			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
DG-Net [27]	CVPR 2019	94.8	-	-	86.0	86.6	-	-	74.8
MpRL [24]	TIP 2019	87.96	-	-	81.18	81.28	-	-	74.54
CAD-Net [41]	ICCV 2019	83.7	92.7	95.8	-	75.6	86.7	89.6	-
FAPL [25]	TMM 2019	86.07	-	-	77.64	79.04	-	-	70.74
FD-GAN [39]	NeurIPS 2018	90.5	-	-	77.7	80.0	-	-	64.5
CamStyle [40]	CVPR 2018	89.49	-	-	71.55	78.32	-	-	57.61
SL [37]	ECCV 2018	89.43	-	-	72.58	73.58	-	88.75	53.20
Pose-transfer [38]	CVPR 2018	87.65	-	-	68.92	78.52	-	-	56.91
LSRO [26]	ICCV 2017	83.97	-	-	66.07	67.68	-	-	47.13
baseline		94.73	97.92	98.75	86.11	86.52	94.42	96.35	74.69
Ours		**97.62**	**98.52**	**99.10**	**89.63**	**90.88**	**95.08**	**96.93**	**80.41**

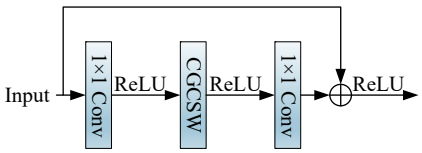


Figure 5: The diagram of the proposed powerful feature representation bottleneck block architecture.

4.1. Datasets and evaluation metrics

In order to verify the effectiveness of our proposed model, we conduct a series of experiments on three well-known and representative person Re-ID datasets: Market-1501, DukeMTMC-reID and MSMT17. The details of each dataset and evaluation metrics are as follows:

Market-1501 includes 36,036 person images of 1,501 identities captured by 6 cameras. The dataset is divided into three parts: 12,936 training images of 751 individuals, 19,732 testing images of 750 individuals and 3,368 query images.

DukeMTMC-reID contains 36,411 person images of 1,812 identities from 8 camera views. Those images are split into three subsets including the training set (16,522 images, 702 identities), the gallery set (17,661 images, 1,100 identities (containing 408 distractor identities)) and the query set (2,228 images, 702 identities).

MSMT17 is composed of 126,411 person images of 4,101 identities from 15 camera angles of view. There are 32,621 training images of 1,041 identities, 82,161 testing images of 3,060 identities and 11,659 query images.

Evaluation metrics. The Cumulative Matching Characteristic (CMC) curve at Rank-1 accuracy, Rank-5 accuracy and Rank-10 accuracy and the mean Average Precision (mAP) are adopted as the evaluation metrics to determine the performance of different person Re-ID models. The Rank-i accuracy represents the ratio of one or more exactly matched images appearing in the top-i ranked images. The mAP value reports the overall accuracy and recall rate, thereby offering a fairly comprehensive evaluation protocol.

4.2. Implementation details

In this paper, the framework of DG-Net [27] is adopted as the baseline. This research uses ResNet-50 [14] pre-trained on ImageNet [48] as our backbone, and only removes the last global average pooling layer and the last fully connected layer (replaced by a fully connected layer with the corresponding number of identities, i.e., 751 for the Market-1501 dataset, 702 for the DukeMTMC-reID dataset and 1041 for the MSMT17 dataset). All the input images are resized to 256×128 during training. The batchsize and the epoch are set to 8 and 100 respectively. The reduction ratio ρ of the CGCSW module in Eq. (5) is experimentally chosen as 8 and is discussed in detail in Section 4.5.2.

In order to further learn powerful feature representations of persons, this research proposes the strong feature representation (SFR) network based on the reported CGCSW module. The SFR network is partly inspired from ResNet-50 [14], which is widely used in computer vision tasks such as classification, recognition, detection, segmentation, retrieval, etc. Nevertheless, unlike the ResNet-50, we do not conduct the 3×3 convolution operation in all blocks of the third and fourth stages, instead, we employ the proposed CGCSW module which could simultaneously focus on the channel and spatial information described in Section 3.1. This replaces the 3×3 convolution to obtain powerful feature representations of individuals for ReID.

Specifically, Fig. 5 illustrates our bottleneck block after embedding the CGCSW module to replace the 3×3 convolution in the corresponding bottleneck block of ResNet-50. This bottleneck block could develop remote channel dependencies and extract spatial information at a finer-grained level. Meanwhile, similar to ResNet-50, a strong feature representation (SFR) network is presented by embedding the bottleneck block in the third and fourth stages to replace the corresponding stages of ResNet-50. The first and second stages of our SFR network have the same architecture as the corresponding branches of ResNet-50, i.e., the first and second stages have the same three and four blocks, respectively. The proposed SFR network inherits the superiority of the CGCSW module and can establish long-term channel attention dependencies and realize the information interaction among channel attention. Furthermore, based on the CGCSW module, our SFR network has significant spatial information mining capabilities. In all experiments of this study, the SFR network will also be pre-trained on ImageNet. It should be noted that we do not embed the bottleneck block into the first and second stages of the original ResNet-50. The detailed explanation could be found in the ablation experiments in Section 4.5.2.

For the FEBS module, the hyperparameters ε_1 and ε_2 in Eq. (12) are preferentially chosen as 0.5 and 0.4, respectively. More detailed explanations of the reason for this choice can be found in Section 4.5.3.

4.3. Comparison with state-of-the-art GAN-based methods

For performance evaluationm we compare our approach with nine state-of-the-art GAN-based methods on two common and representative person Re-ID datasets, i.e., the Market-1501 and DukeMTMC-reID. Table 1 shows comparative results, indicating that our proposed approach is superior to all state-of-the-art GAN-based algorithms with clear advantages on both the Market-1501 and DukeMTMC-reID datasets. More specifically, the proposed method surpasses the second-best model DG-Net by +2.82% (97.62 - 94.8)/+3.63% (89.63 - 86.0) in Rank-1 accuracy/mAP using Market-1501 and +4.28% (90.88 - 86.6)/+5.61% (80.41 - 74.8) in Rank-1 accuracy/mAP using DukeMTMC-reID. In addition, compared with CAD-Net which has the second-best results in Rank-5 accuracy and Rank-10 accuracy, our approach achieves the best improvement of +5.82% (98.52 - 92.7)/+3.30% (99.10 - 95.8) in Rank-5 accuracy/Rank-10 accuracy using Market-1501 and +8.38% (95.08 - 86.7)/+7.33% (96.93 - 89.6) in Rank-5 accuracy/Rank-10 accuracy using DukeMTMC-reID. It can be seen that the experimental results are not inferior to the other GAN-based methods listed in Table 1. Therefore, these results demonstrate the effectiveness of the proposed AE-Net learning framework in comparison to other approaches.

Table 2: Comparison of experimental results with state-of-the-art methods using the Market-1501 dataset. Rank-1, Rank-5, Rank-10 and mAP are listed. Among these results, the first-, second- and third-best results are highlighted in red, green and blue respectively.

Methods	References	Market-1501			
		Rank-1	Rank-5	Rank-10	mAP
BV [49]	ICCV 2021	96.0	-	-	89.2
ES-Net [50]	TIP 2021	95.7	-	-	88.6
PAT [51]	CVPR 2021	95.4	-	-	88.0
TransReID [52]	ICCV 2021	95.2	-	-	89.5
CDNet [53]	CVPR 2021	95.1	-	-	86.0
ASSP [54]	CVPR 2021	95.0	-	-	87.3
ADC [55]	CVPR 2021	94.8	97.2	98.0	87.7
PCB+RPP [56]	TPAMI 2021	93.8	97.5	98.5	81.6
OAMN [57]	ICCV 2021	93.2	-	-	79.8
KPM-GSRW [58]	TPAMI 2021	93.1	97.2	98.0	84.7
RGA-SC [59]	CVPR 2020	96.1	-	-	88.4
PISNet [60]	ECCV 2020	95.6	-	-	87.1
MuDeep [5]	TPAMI 2020	95.34	-	-	84.66
ISP [61]	ECCV 2020	95.3	98.6	-	88.6
GASM [62]	ECCV 2020	95.3	-	-	84.7
ICT+CE [63]	TIP 2020	94.4	-	-	84.9
CBN [64]	ECCV 2020	94.3	97.9	98.7	83.6
DSA [65]	CVPR 2019	95.7	-	-	87.6
BDB+Cut [33]	ICCV 2019	95.3	-	-	86.7
OSNet [66]	ICCV 2019	94.8	-	-	84.9
CASN [32]	CVPR 2019	94.4	-	-	82.8
PIE+*Siam.* [13]	TIP 2019	89.06	96.01	97.13	70.69
Ours		97.62	98.52	99.10	89.63

Table 3: Comparison of experimental results with state-of-the-art methods using the DukeMTMC-reID dataset. Rank-1, Rank-5, Rank-10 and mAP are listed. Among these results, the first-, second- and third-best results are highlighted in red, green and blue respectively.

Methods	References	DukeMTMC-reID			
		Rank-1	Rank-5	Rank-10	mAP
TransReID [52]	ICCV 2021	90.7	-	-	82.6
BV [49]	ICCV 2021	90.5	-	-	80.6
ES-Net [50]	TIP 2021	89.2	-	-	78.8
PAT [51]	CVPR 2021	88.8	-	-	78.2
CDNet [53]	CVPR 2021	88.6	-	-	76.8
ASSP [54]	CVPR 2021	88.2	-	-	76.1
ADC [55]	CVPR 2021	87.4	92.1	95.5	74.9
OAMN [57]	ICCV 2021	86.3	-	-	72.6
PCB+RPP [56]	TPAMI 2021	84.5	-	-	71.5
KPM-GSRW [58]	TPAMI 2021	83.4	90.0	91.7	71.3
ISP [61]	ECCV 2020	89.6	95.5	-	80.0
PISNet [60]	ECCV 2020	88.8	-	-	78.7
GASM [62]	ECCV 2020	88.3	-	-	74.4
MuDeep [5]	TPAMI 2020	88.19	-	-	75.63
ICT+CE [63]	TIP 2020	88.1	-	-	75.3
CBN [64]	ECCV 2020	84.8	92.5	95.2	70.1
BDB+Cut [33]	ICCV 2019	89.0	-	-	76.0
OSNet [66]	ICCV 2019	88.6	-	-	73.5
CASN [32]	CVPR 2019	87.7	-	-	73.7
DSA [65]	CVPR 2019	86.2	-	-	74.3
PIE [13]	TIP 2019	80.84	-	-	64.09
Ours		90.88	95.08	96.93	80.41

Table 4: Comparison of experimental results with state-of-the-art methods using the MSMT17 dataset. Rank-1, Rank-5, Rank-10 and mAP are listed. Among these results, the first-, second- and third-best results are highlighted in red, green and blue respectively.

Methods	References	Rank-1	Rank-5	Rank-10	mAP
ES-Net [50]	TIP 2021	80.9	-	-	57.3
CDNet [53]	CVPR 2021	78.9	-	-	54.7
PCB+RPP [56]	TPAMI 2021	69.8	83.3	86.7	43.6
KPM-GSRW [58]	TPAMI 2021	71.8	83.3	87.0	47.8
RGA-SC [59]	CVPR 2020	80.3	-	-	57.5
GASM [62]	ECCV 2020	79.5	-	-	52.5
OSNet [66]	ICCV 2019	78.7	-	-	52.9
DG-Net [27]	CVPR 2019	77.2	87.4	90.5	52.3
Ours		82.09	90.11	92.57	57.82

4.4. Comparison with Other State-of-the-art Methods

This study evaluates the proposed approach against a number of state-of-the-art approaches on three well-known and representative person Re-ID datasets including the Market-1501, DukeMTMC-reID and MSMT17.

4.4.1. Evaluation on the Market-1501 dataset

This research compares our method with 22 state-of-the-art approaches using the Market-1501 dataset. The Market-1501 is a large-scale person Re-ID dataset with sufficient training and testing images, which is especially suitable for data-driven deep learning approaches. As can be seen in Table 2, our method surpasses all the compared state-of-the-art approaches with increases in Rank-1 accuracy and slight increases both in Rank-10 accuracy and mAP. Specifically, our approach achieves improvements of +1.52% (97.62 - 96.1)/+0.40% (99.10 - 98.7)/+0.13% (89.63 - 89.5) over the second-best RGA-SC/CBN/TransReID in Rank-1 accuracy/Rank-10 accuracy/mAP. Furthermore, the proposed method performs well (ranking second) in Rank-5 accuracy, only +0.08% (98.6 - 98.52) lower than the best technique, ISP. It can be seen that the experimental results of this study are superior to the other approaches. This proves the efficacy of the proposed CGCSW module, the FEBS module and the stage-wise consistency loss.

4.4.2. Evaluation using the DukeMTMC-reID dataset

This study evaluates the proposed approach against 21 state-of-the-art methods using the DukeMTMC-reID dataset. Compared with Market-1501, DukeMTMC-reID has more surveillance camera views and complex scenes, which leads to significant changes in the background and image resolution, making the person Re-ID task more challenging. The comparative results shown in Table 3 indicate that the proposed approach is superior to all the state-of-the-art methods with clear improvements in Rank-10 accuracy and slight improvement in Rank-1 accuracy. More specifically, our method achieves improvements of +1.43% (96.93 - 95.5)/+0.18% (90.88 - 90.7) over the second-best approaches, ADC/TransReID, in Rank-10 accuracy/Rank-1 accuracy. We observe that the proposed approach achieves the same result compared with the best technique, ISP considering Rank-5 accuracy, surpassing the third-best CBN by an improvement of +2.58% (95.08 - 92.5). In addition, the proposed method performs well (ranking third) in mAP, only +0.19% (80.6 - 80.41) lower than the second-best BV. As shown in Table 3, the proposed approach presented in this research has the best performance compared with others similar techniques.

Table 5: Ablation Experimental results of the proposed AE network both on the Market-1501 and DukeMTMC-reID datasets. Rank-1, Rank-5, Rank-10 and mAP are listed.

Methods	Market-1501				DukeMTMC-reID			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
baseline	94.73	97.92	98.75	86.11	86.52	94.42	96.35	74.69
SFR	96.85	98.13	98.94	88,26	89.63	94.87	96.66	78.52

Table 6: Parameter sizes (millions, denoted M) of the CGCSW module in the AE network and the performance at different reduction ratios on the Market-1501 dataset.

Ratio ρ	Parameters	Rank-1	Rank-5	Rank-10	mAP
2	1.19M	96.94	98.09	99.02	88.17
4	0.60M	96.81	98.11	98.90	88.21
8	0.31M	96.85	98.13	98.94	88.26
16	0.16M	96.76	98.03	98.81	88.19

4.4.3. Evaluation on the MSMT17 dataset

This paper compares the proposed method with eight state-of-the-art approaches on the MSMT17 dataset. The MSMT17 consists of more person images than the other two datasets mentioned above, making it closer to a real situation and thus enabling further validation of the algorithm's ability to recognize real-world scenes. Table 4 shows the comparative results, indicating that the proposed method surpasses all the compared approaches. Specifically, the proposed algorithm surpasses the second-best techniques, ES-Net/DG-Net, by +1.19% (82.09 - 80.9)/ +2.71% (90.11 - 87.4) (and +2.07% (92.57 - 90.5)) in Rank-1 accuracy/Rank-5 accuracy (and Rank-10 accuracy). Additionally, we achieve an advantage in mAP, surpassing the second-best technique RGA-SC by +0.32% (57.82 - 57.5). The experimental results presented in Table 4 on the MSMT17 dataset verify the superiority of the proposed AE-Net.

4.5. Ablation analysis

In this section, this research conducts ablation analysis on the Market-1501 and DukeMTMC-reID datasets. We analyze the efficacy of each component of the proposed appearance-enriched neural network with foreground enhancement (AE-Net), including the strong feature representation (SFR) network, the channel-grouped convolution and squeeze weighted (CGCSW) module along with the foreground-enhanced and background-suppressed (FEBS) module.

4.5.1. Effectiveness of the strong feature representation network

We validate the efficacy of the strong feature representation (SFR) network using ablation experiments. As can be seen in Table 5, the proposed SFR network outperforms the baseline with significant advantages both on the Market-1501 and DukeMTMC-reID datasets. More specifically, the SFR network achieves obvious improvements of +2.12% (96.85 - 94.73)/+2.15% (88.26 - 86.11) over the baseline in Rank-1 accuracy and mAP using the Market-1501, and +3.11% (89.63 - 86.52)/+3.83% (78.52 - 74.69) in Rank-1 accuracy and mAP using the DukeMTMC-reID. This indicates that our SFR network greatly enhances the feature representation, i.e., the SFR network could learn more detailed and distinctive features of the individual representation than the baseline, thereby further providing useful features of individuals for the FEBS module.

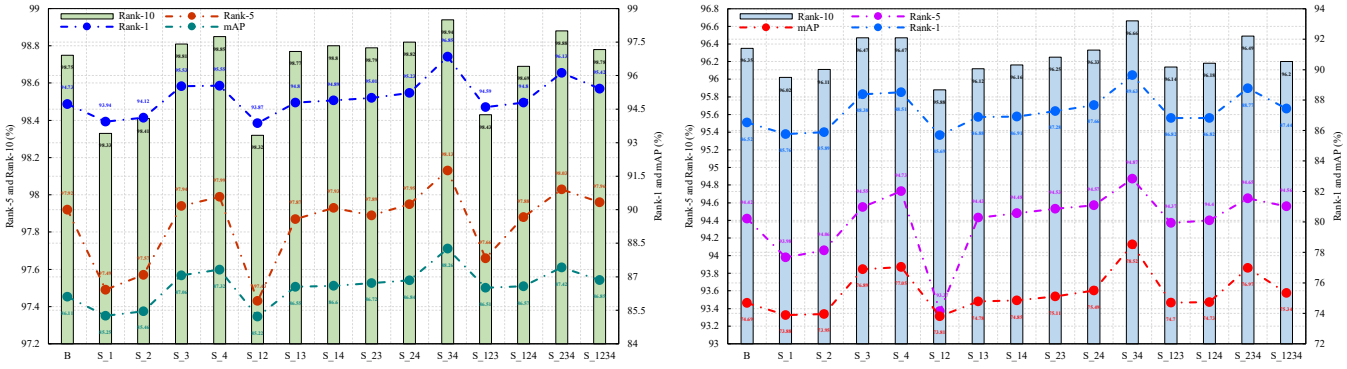
4.5.2. Effectiveness of the proposed CGCSW module

This study discusses the impact of different combinations of the proposed CGCSW module on the person Re-ID task. Fig. 6 shows 14 ablation results using both the Market-1501 and DukeMTMC-reID datasets. The abbreviation S_1 represents that the CGCSW module is only embedded in the first stage of the ResNet-50, and the other abbreviations are similar to the S_1 where the corresponding number indicates the stage where CGCSW is embedded, i.e. S_1234 indicates embedding in stages 1, 2, 3 and 4. Specifically, the results of the S_1, S_2 and S_12 are lower than the baseline and the other listed combinations of the CGCSW module, which suggests that our CGCSW module is not suitable for embedding into the first two stages of ResNet-50. Among these various combinations of S_1, S_2, S_3 and S_4, the S_34 obtains the best results compared with the baseline and the others. This clarifies the reason why we adopt this combination as the final SFR network.

In addition, the value of the reduction ratio ρ reported in Eq. (5) affects performance and computational costs of the CGCSW module. This research conducts comparative experiments using the Market-1501 dataset to obtain the best value of ρ. It can be seen in Table 6 that the performance does not improve monotonically with the increase of ρ, especially when the value of ρ is set to 8, a good trade-off between the performance and computation is achieved. Therefore, we adopt $\rho = 8$ as the reduction ratio used in all experiments.

4.5.3. Effectiveness of the proposed FEBS module

In order to illustrate the efficacy of our FEBS module, we conduct further ablation experiments: (1) adding the FEBS module to the baseline and the SFR network respectively (shown in Table 7); (2) the analysis of the two hyper-parameters ε_1 and ε_2 on the Market-1501 dataset (shown in Fig. 7). More specifically, the results in Table 7 show that the baseline+FEBS and the SFR+FEBS outperform the baseline and the SFR network respectively with significant advantages both in Rank-1 accuracy and mAP. This suggests significant foreground-enhanced and background-suppressed capability of the FEBS module proposed in this paper. Additionally, this study analyzes the effect of the hyperparameters ε_1 and ε_2 of the FEBS module. Lots of comparative experiments have been conducted using the Market1501 dataset, and the results report that the accuracy (Rank-1, Rank-5, Rank-10 and mAP) reaches its highest



(a) Market-1501 (b) DukeMTMC-reID

Figure 6: Histogram and line chart of the ablation experimental results for the proposed CGCSW module using the Market-1501 and DukeMTMC-reID datasets.

when $\varepsilon_1 = 0.5$ and $\varepsilon_2 = 0.4$. We found the optimal combination of ε_1 and ε_2 in the range of 0 to 0.9 with a step length of 0.1. As can be seen in Fig. 7, the accuracy of Rank-1, Rank-5, Rank-10 and mAP are obviously lower than our best combination ($\varepsilon_1 = 0.5, \varepsilon_2 = 0.4$) when $\varepsilon_1 = \varepsilon_2 = 0$ which means that the module ignores the influence of the bright and complex backgrounds in the person Re-ID task. These experimental results fully validate the significance role of the foreground-enhanced and background-suppressed feature module and the suitability of the hyperparameter selection.

4.6. Evaluation of the generated samples

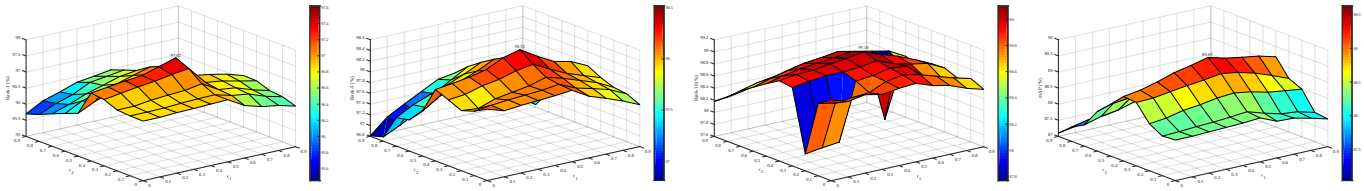
In this section, we present the evaluation of the samples generated by the GAN-based approaches from both a quantitative and qualitative perspective.

Quantitative evaluation. This paper employs the Structural Similarity (SSIM) [67] and Fréchet Inception Distance (FID) [68] to evaluate the quality of the generated samples. SSIM measures the structural similarity between two images. FID measures the distance between the feature vectors of a real image and a generated image, which means that a lower score has a high correlation with a higher quality image. Table 8 reports the comparative results using the Market-1501 dataset. Specifically, the comparative results are superior to the other methods listed in Table 8. These results indicate the high quality of the generated samples, and thus could help to enhance the overall person Re-ID performance.

Qualitative evaluation. Fig. 8 shows examples of real and generated images on the Market-1501 and DukeMTMC-reID datasets, which illustrate the capacity of the proposed network. This study presents visualized examples to evaluate the appearance-enriched samples qualitatively. Fig. 9 presents visual comparison results for the same samples from the real, DG-Net and our approach, respectively. It can be seen from Fig. 9 that the samples generated by DG-Net contain artifacts and visual blurs (such as hair, hip, shoulder, chest, face, etc.), while our generated samples look more natural. These visual comparison results suggest that our AE-Net could learn more detailed and distinctive features of the individuals structure and appear-

Table 7: Ablation experimental results for the proposed FEBS module using the Market-1501 and DukeMTMC-reID datasets. Rank-1, Rank-5, Rank-10 and mAP are listed.

Methods	Market-1501				DukeMTMC-reID			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
baseline	94.73	97.92	98.75	86.11	86.52	94.42	96.35	74.69
baseline+FEBS	95.82	98.01	98.89	87.76	87.90	94.63	96.48	76.30
SFR	96.85	98.13	98.94	88,26	89.63	94.87	96.66	78.52
SFR+FEBS	97.62	98.52	99.10	89.63	90.88	95.08	96.93	80.41



(a) Rank-1 (b) Rank-5 (c) Rank-10 (d) mAP

Figure 7: Three-dimensional surface plots of ablation results for different combinations of the ε_1 and ε_2 using the Market-1501 dataset. Rank-1, Rank-5, Rank-10 and mAP are listed.

Table 8: SSIM and FID results to evaluate the generated samples by the FA-Net on the Market-1501 dataset. The best results among these methods are highlighted in bold.

Methods	References	SSIM	FID
DG-Net [27]	CVPR 2019	0.360	18.24
SL [37]	ECCV 2018	0.335	54.23
FD-GAN [39]	NeurIPS 2018	0.247	257.00
PG^2-GAN [69]	NeurIPS 2017	–	151.16
LSGAN [70]	ICCV 2017	–	136.26
Real		0.350	7.22
baseline		0.355	18.72
Ours		**0.368**	**16.19**

ance than the DG-Net, thereby further assisting in improving the performance of the person Re-ID task.

5. Conclusions

This paper introduces how the CGCSW module and the FEBS module can be combined to strengthen the recognition capacity of our proposed AE-Net. We employ the CGCSW module to explore the useful feature representation of persons. The FEBS module is reported to enhance the foreground of person images and weaken the influence of the background simultaneously. Furthermore, the stage-wise consistency loss enables our model to be consistent between the foreground-enhanced and background-suppressed stages. This research verifies all the vital components of the presented framework and performs comprehensive experiments on three well-known and representative datasets. The experimental results indicate the capability of the proposed method, which is superior to many state-of-the-art approaches for the task of person Re-ID.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Figure 8: Examples of the real and generated images using the Market-1501 (the first two lines) and DukeMTMC-reID (the bottom two lines) datasets.

Figure 9: Visual comparison results among the same samples from the real, DG-Net and proposed approach, respectively.

Acknowledgments

The research is supported by National Natural Science Foundation of China (No. 61973066, 61471110) , Major Science and Technology Projects of Liaoning Province (No. 2021JH1/10400049), Fundation of Key Laboratory of Aerospace System Simulation (No. 6142002200301), Fundation of Key Laboratory of Equipment Reliability (No. WD2C20205500306), Open Research Projects of Zhejiang Lab (No. 2019KD0AD01/006) and Major Science and Technology Innovation Engineering Projects of Shandong Province (No. 2019JZZY010128).

References

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. Hoi, Deep learning for person re-identification: A survey and outlook, IEEE TPAMI (2021) 1–1.
[2] H. Tian, X. Zhang, L. Lan, Z. Luo, Person re-identification via adaptive verification loss, Neurocomputing 359 (2019) 93–101.
[3] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, Interaction-and-aggregation network for person re-identification, in: CVPR, 2019, pp. 9317–9326.
[4] Z. Chang, Z. Qin, H. Fan, H. Su, H. Yang, S. Zheng, H. Ling, Weighted bilinear coding over salient body parts for person re-identification, Neurocomputing 407 (2020) 454–464.
[5] X. Qian, Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, Leader-based multi-scale attention deep architecture for person re-identification, IEEE TPAIM 42 (2) (2020) 371–385.
[6] H. Liu, Z. Xiao, B. Fan, H. Zeng, Y. Zhang, G. Jiang, PrGCN: Probability prediction with graph convolutional network for person re-identification, Neurocomputing 423 (2021) 57–70.
[7] Y. Liu, Y. Zhang, S. Coleman, B. Bhanu, S. Liu, A new patch selection method based on parsing and saliency detection for person re-identification, Neurocomputing 374 (2020) 86–99.
[8] M. Jiang, B. Leng, G. Song, Z. Meng, Weighted triple-sequence loss for video-based person re-identification, Neurocomputing 381 (2020) 314–321.
[9] J. Peng, H. Wang, F. Xu, X. Fu, Cross domain knowledge learning with dual-branch adversarial network for vehicle re-identification, Neurocomputing 401 (2020) 133–144.
[10] S. Liao, Y. Hu, X. Zhu, S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: CVPR, 2015, pp. 2197–2206.
[11] R. Zhao, W. Oyang, X. Wang, Person re-identification by saliency learning, TPAMI 39 (2) (2016) 356–370.
[12] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: CVPR, 2017, pp. 384–393.
[13] L. Zheng, Y. Huang, H. Lu, Y. Yang, Pose-invariant embedding for deep person re-identification, IEEE TIP 28 (9) (2019) 4500–4509.

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.

[15] W. Chen, X. Chen, J. Zhang, K. Huang, A multi-task deep network for person re-identification, in: AAAI, 2017, pp. 3988–3994.

[16] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person reidentification, ACM TOMM 14 (1) (2017) 1–20.

[17] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, X. Xue, Multi-scale deep learning architectures for person re-identification, in: ICCV, 2017, pp. 5399–5408.

[18] C. Wang, Q. Zhang, C. Huang, W. Liu, X. Wang, Mancs: A multi-task attentional network with curriculum sampling for person re-identification, in: ECCV, 2018, pp. 365–381.

[19] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, Y. Yang, Salience-guided cascaded suppression network for person re-identification, in: CVPR 2020, 2020, pp. 3300–3310.

[20] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, in: CVPR, 2018, pp. 2285–2294.

[21] C. Song, Y. Huang, W. Ouyang, L. Wang, Mask-guided contrastive attention model for person re-identification, in: CVPR, 2018, pp. 1179–1188.

[22] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, M. Shah, Human semantic parsing for person re-identification, in: CVPR, 2018, pp. 1062–1071.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: NeurIPS, 2014, pp. 2672–2680.

[24] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, J. Zhang, Multi-pseudo regularized label for generated data in person re-identification, IEEE TIP 28 (3) (2019) 1391–1403.

[25] G. Ding, S. Zhang, S. Khan, Z. Tang, J. Zhang, F. Porikli, Feature affinity-based pseudo labeling for semi-supervised person re-identification, IEEE TMM 21 (11) (2019) 2891–2902.

[26] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: ICCV, 2017, pp. 3754–3762.

[27] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, J. Kautz, Joint discriminative and generative learning for person re-identification, in: CVPR, 2019, pp. 2138–2147.

[28] D. Chen, S. Zhang, W. Ouyang, J. Yang, Y. Tai, Person search via a mask-guided two-stream cnn model, in: ECCV, 2018, pp. 734–750.

[29] H. Cai, Z. Wang, J. Cheng, Multi-scale body-part mask guided attention for person re-identification, in: CVPR Workshops, 2019, pp. 0–0.

[30] D. Chen, S. Zhang, W. Ouyang, J. Yang, Y. Tai, Person search by separated modeling and a mask-guided two-stream cnn model, IEEE TIP 29 (2020) 4669–4682.

[31] L. Bao, B. Ma, H. Chang, X. Chen, Masked graph attention network for person re-identification, in: CVPR Workshops, 2019, pp. 0–0.

[32] M. Zheng, S. Karanam, Z. Wu, R. J. Radke, Re-identification with consistent attentive siamese networks, in: CVPR, 2019, pp. 5735–5744.

[33] Z. Dai, M. Chen, X. Gu, S. Zhu, P. Tan, Batch dropblock network for person re-identification and beyond, in: ICCV, 2019, pp. 3691–3701.

[34] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, S. Zhang, Towards rich feature discovery with class activation maps augmentation for person re-identification, in: CVPR, 2019, pp. 1389–1398.

[35] B. Chen, W. Deng, J. Hu, Mixed high-order attention network for person re-identification, in: ICCV, 2019, pp. 371–381.

[36] W. Li, X. Zhu, S. Gong, Scalable person re-identification by harmonious attention, IJCV 128 (6) (2020) 1635–1653.

[37] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, X. Xue, Pose-normalized image generation for person re-identification, in: ECCV, 2018, pp. 650–667.

[38] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, J. Hu, Pose transferrable person re-identification, in: CVPR, 2018, pp. 4099–4108.

[39] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, H. Li, FD-GAN: Pose-guided feature distilling gan for robust person re-identification, in: NeurIPS, 2018, pp. 1230–1241.

[40] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camera style adaptation for person re-identification, in: CVPR, 2018, pp. 5157–5166.

[41] Y.-J. Li, Y.-C. Chen, Y.-Y. Lin, X. Du, Y.-C. F. Wang, Recover and identify: A generative dual model for cross-resolution person re-identification, in: ICCV, 2019, pp. 8090–8099.

[42] H. Zhang, K. Zu, J. Lu, Y. Zou, D. Meng, EPSANet: An efficient pyramid split attention block on convolutional neural network, arXiv preprint arXiv:2105.14447 (2021).

[43] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, CBAM: Convolutional block attention module, in: ECCV, 2018, pp. 3–19.

[44] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: CVPR, 2018, pp. 7132–7141.

[45] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in: ICML, 2010, pp. 807–814.

[46] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: ICCV, 2015, pp. 1116–1124.

[47] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: CVPR, 2018, pp. 79–88.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: CVPR, Ieee, 2009, pp. 248–255.

[49] C. Yan, G. Pang, L. Wang, J. Jiao, X. Feng, C. Shen, J. Li, BV-Person: A large-scale dataset for bird-view person re-identification, in: ICCV, 2021, pp. 10943–10952.

[50] D. Shen, S. Zhao, J. Hu, H. Feng, D. Cai, X. He, ES-Net: Erasing salient parts to learn more in re-identification, IEEE TIP 30 (2021) 1676–1686.

[51] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, F. Wu, Diverse part discovery: Occluded person re-identification with part-aware transformer, in: CVPR, 2021, pp. 2898–2907.

[52] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, TransReID: Transformer-based object re-identification, in: ICCV, 2021, pp. 15013–15022.

[53] H. Li, G. Wu, W.-S. Zheng, Combined depth space based architecture search for person re-identification, in: CVPR, 2021, pp. 6729–6738.

[54] J. Chen, X. Jiang, F. Wang, J. Zhang, F. Zheng, X. Sun, W.-S. Zheng, Learning 3d shape feature for texture-insensitive person re-identification, in: CVPR, 2021, pp. 8146–8155.

[55] A. Zhang, Y. Gao, Y. Niu, W. Liu, Y. Zhou, Coarse-to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck, in: CVPR, 2021, pp. 598–607.

[56] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, S. Wang, Learning part-based convolutional features for person re-identification, IEEE TPAMI 43 (3) (2021) 902–917.

[57] P. Chen, W. Liu, P. Dai, J. Liu, Q. Ye, M. Xu, Q. Chen, R. Ji, Occlude them all: Occlusion-aware attention network for occluded person re-id, in: ICCV, 2021, pp. 11833–11842.

[58] Y. Shen, T. Xiao, S. Yi, D. Chen, X. Wang, H. Li, Person re-identification with deep kronecker-product matching and group-shuffling random walk, IEEE TPAMI 43 (5) (2021) 1649–1665.

[59] Z. Zhang, C. Lan, W. Zeng, X. Jin, Z. Chen, Relation-aware global attention for person re-identification, in: CVPR, 2020, pp. 3186–3195.

[60] S. Zhao, C. Gao, J. Zhang, H. Cheng, C. Han, X. Jiang, X. Guo, W.-S. Zheng, N. Sang, X. Sun, Do not disturb me: Person re-identification under the interference of other pedestrians, in: ECCV, Springer, 2020, pp. 647–663.

[61] K. Zhu, H. Guo, Z. Liu, M. Tang, J. Wang, Identity-guided human semantic parsing for person re-identification, in: ECCV, Springer, 2020, pp. 346–363.

[62] L. He, W. Liu, Guided saliency feature learning for person re-identification in crowded scenes, in: ECCV, Springer, 2020, pp. 357–373.

[63] F. Xu, B. Ma, H. Chang, S. Shan, Isosceles constraints for person re-identification, IEEE TIP 29 (2020) 8930–8943.

[64] Z. Zhuang, L. Wei, L. Xie, T. Zhang, H. Zhang, H. Wu, H. Ai, Q. Tian, Rethinking the distribution gap of person re-identification with camera-based batch normalization, in: ECCV, Springer, 2020, pp. 140–157.

[65] Z. Zhang, C. Lan, W. Zeng, Z. Chen, Densely semantically aligned person re-identification, in: CVPR, 2019, pp. 667–676.

[66] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Omni-scale feature learning for person re-identification, in: ICCV, 2019, pp. 3702–3712.

[67] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE TIP 13 (4) (2004) 600–612.

[68] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: NeurIPS, 2017, pp. 6629–6640.

[69] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool, Pose guided person image generation, in: NeurIPS, 2017, pp. 405–415.

[70] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, S. Paul Smolley, Least squares

generative adversarial networks, in: ICCV, 2017, pp. 2794–2802.