Salience-Guided Cascaded Suppression Network for Person Re-identification

Xuesong Chen * [1, 2] Canmiao Fu [3] Yong Zhao [1]
Feng Zheng † [2] Jingkuan Song [6] Rongrong Ji [4, 7] Yi Yang [5]
[1]School of Electronic and Computer Engineering, Peking University, China
[2]Southern University of Science and Technology, China [3]Tencent, China [4]Xiamen University, China
[5]The ReLER Lab, Centre for AI, University of Technology Sydney, Australia
[6]University of Electronic Science and Technology of China, China [7]Peng Cheng Laboratory, China
{cedarchen,fcm,yongzhao}@pku.edu.cn, zhengf@sustech.edu.cn

Abstract

Employing attention mechanisms to model both global and local features as a final pedestrian representation has become a trend for person re-identification (Re-ID) algorithms. A potential limitation of these methods is that they focus on the most salient features, but the re-identification of a person may rely on diverse clues masked by the most salient features in different situations, e.g., body, clothes or even shoes. To handle this limitation, we propose a novel Salience-guided Cascaded Suppression Network (SCSN) which enables the model to mine diverse salient features and integrate these features into the final representation by a cascaded manner.

Our work makes the following contributions: (i) We observe that the previously learned salient features may hinder the network from learning other important information. To tackle this limitation, we introduce a cascaded suppression strategy, which enables the network to mine diverse potential useful features that be masked by the other salient features stage-by-stage and each stage integrates different feature embedding for the last discriminative pedestrian representation. (ii) We propose a Salient Feature Extraction (SFE) unit, which can suppress the salient features learned in the previous cascaded stage and then adaptively extracts other potential salient feature to obtain different clues of pedestrians. (iii) We develop an efficient feature aggregation strategy that fully increases the network's capacity for all potential salience features. Finally, experimental results demonstrate that our proposed method outperforms the state-of-the-art methods on four large-scale datasets. Especially, our approach exceeds the current best method by over 7% on the CUHK03 dataset.

*This work was done when Xuesong Chen visited to Feng Zheng Lab at SUSTech.

†Corresponding author.

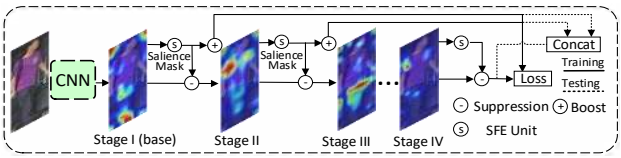

Figure 1. The insight of the Salience-guided Cascaded Suppression Network (SCSN). For training, each stage is guided by the gradient from the loss function. During testing, different stage's features will be concatenated to generate the final diverse pedestrian representation. Benefiting from the suppression strategy, potential important features can stand out in the next stage, which enables different stages to discover diverse clues of pedestrians.

1. Introduction

Given a probe person image, the Person Re-identification (Re-ID) task aims to search the picture that most likely belongs to the same pedestrian from the gallery (the candidate picture set). It is commonly applied to address the issues of cross-camera tracking and surveillance security and can be considered as an image retrieval problem.

To this end, most existing Re-ID methods focus on learning discriminative and robust features to match the pair of images in response to various challenges, including varying viewing angles, lighting intensity and body pose variations. Specifically, recent studies [9, 12, 27, 32] have shown that combining part-based local features with global features is an effective strategy to enhance the feature representation. In general, considering a whole image, the global feature is robust to the appearance changes and spatial location variations. However, lacking the supervision of fine-grained characterizations, global features may focus on interference information, such as backgrounds, which is not expected. Besides, global features are prone to ignore the information of some small regions which can make contributions for discriminative pedestrian representation. Motivated by such observations, the attention mechanism and part-based

models were introduced to address these issues [5, 7]. The employment of attention mechanisms can enforce the model to capture the discriminative local features of human bodies and reduce the interference of different variations in the background. Meanwhile, part-based models could concentrate on learning more fine-grained local salience features of different human body parts by dividing feature maps into horizontal parts. So, with attention mechanisms, aggregating local features and global features has become a trend for person Re-ID and it has achieved promising results.

Nevertheless, one crucial limitation of these global-local methods, including attention-based and part-based, is the lack of exploration of how to effectively extract discriminative potential salience features of different pedestrians. On the one hand, the attention-based methods [5, 7, 16, 38] mainly focus on the discriminative appearances of the human body. However, the attention mechanism trained in a weakly-supervised manner tends to learn the "easiest" features at a compact subspace due to the partial learning behavior of deep models [3, 4]. In other words, deep models easily focus on surface distribution regularities rather than more general and diverse concepts, so that they are prone to ignore potential information of pedestrians. On the other hand, part-based methods [24, 36, 49] handle misalignment and provide richer fine-grained local features by dividing the input into many horizontal stripes. However, with the number of parts increases, the improvement of accuracy is minor and even gets worse. Because too fine division deprives the semantic information of each part and makes the network redundant as well. Moreover, if all kinds of features are indiscriminately concatenated, some significant discriminant features which are not distinctive in intensity will be masked by other salient features. Therefore, how to efficiently extract diverse salient features and how to integrate these features reasonably are worthy of discussion for the Re-ID task.

In this paper, to further improve the model's feature representation capability, we present a salience-guided cascaded feature suppression mechanism that enables the network to adaptively extract all potential salient pedestrian features. More specifically, we propose a feature aggregation strategy which consists of a Residual Dual Attention Module (RDAM) and a Non-local Multi-stage Feature Fusion (NMFF) block, to better aggregate low-level and high-level features of the backbone, and a Salient Feature Extract (SFE) unit to effectively yet efficiently extract diverse potential features. With the help of the feature aggregation strategy, our network can make better use of low-level features, such as the color and texture of the clothes, which greatly improves the feature representation capability of the backbone. Hence, the cascaded suppression head with SFE units can extract salience features via a cascaded suppression update. In practice, we first employ a global stage on top of the backbone to extract the most salient region-level information with the SFE unit. In order to boost information flow in our feature suppression mechanism, the salient feature learned at a certain stage is first integrated with the global feature to enhance this stage's feature discriminability, and then it will be suppressed to get the salience-free input feature for the next stage. Similarly, for the rest stages, the network will mine some other important potential features with the SFE unit after the previous salient feature being suppressed. We illustrate the salience-guided cascaded suppression network in Fig. 1.

To summarize, our proposed work makes the following contributions:

- We introduce a novel cascaded feature suppression mechanism that can mine all potential salient features stage-by-stage and integrate these discriminative salience features with the global feature, forming the final diverse feature representation of pedestrians.
- We devise a Salient Feature Extraction(SFE) unit to adaptively extract potential salient features by suppressing the most salient features.
- We incorporate an efficient feature aggregation strategy, consisting of the RDAM and the NMFF block, which increases the network's capacity for all potential salience features.
- Extensive experiments on Market1501 [50], DukeMTMC-ReID [52], CUHK03 [22] and MSMT17 [42] demonstrate that our method significantly outperforms the existing state-of-the-art methods on four popular benchmarks.

2. Related Work

Due to the improvement of computing power, in recent years, deep learning based methods are developed to address the person Re-ID task. Below, we review the most representative methods which are related to our work.

Part-based algorithms: Pyramid [49], PCB [36] and MGN [41] achieve state-of-the-art performance by integrating global features and many stripe-based features. This strategy often requires a complex network to learn and combine different levels of features, which bringing performance improvements but suffering from over-fitting and information redundancy. Compared with [36, 41, 49] which divide the feature maps horizontally, other methods learn more semantic local features guided by prior knowledge. Attention-related methods [5, 7, 16, 38] employ different attention modules to improve feature representation and achieve further performance gains on the baseline. Furthermore, incorporating prior knowledge such as the human body structure has also been proved to be an effective method [23, 29, 33, 39, 45, 47]. Meanwhile, GLAD [43]

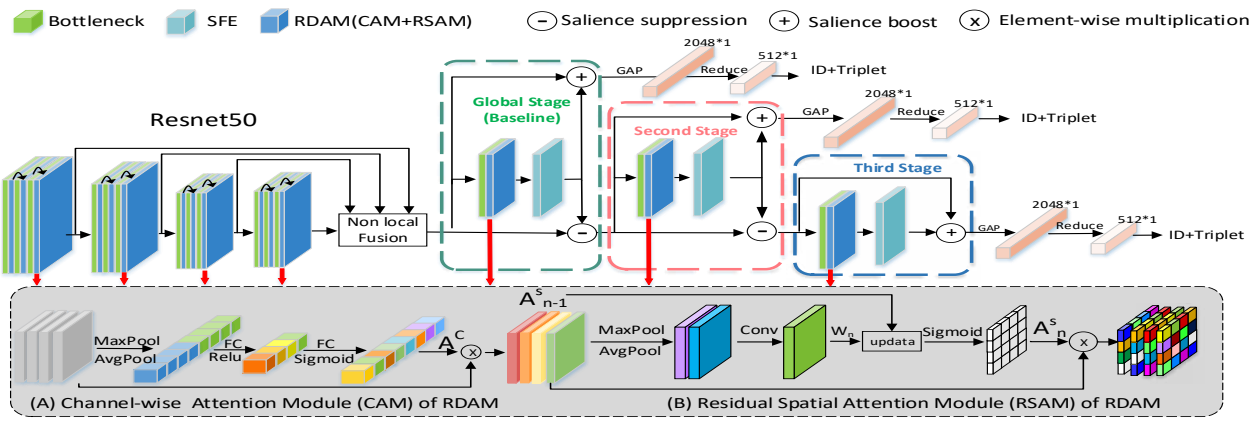

Figure 2. The pipeline of our Salience-guided Cascaded Suppression Network (SCSN). After the modified ResNet50 backbone, we employ several independent feature suppression stages. During training, both the ID loss and Triplet loss are used to supervise the learning of these stages, respectively. In the testing, all features of different stages are concatenated together as the final descriptor of a pedestrian image.

integrates features from both local and global regions using the detected human body parts. Similar to the attention strategies, [48] employs composite models to extract specific salience features from different parts of the human body in an unsupervised manner. However, such methods are prone to possible noises from the pose estimation and semantic parsing algorithms.

Non-part-based algorithms: Recently, to improve the global feature's attention to local information, some adversarial occluded samples [18] and a synthetic dataset [2] are constructed to improve the Re-ID model's feature extraction ability. Moreover, lots of metric learning methods [6, 14, 35] aim to enlarge the inter distance while reducing intra distinction, which improve the representation of Re-ID task. And in our work, we also use the triplet loss [14]. Further, in GSRW [30], a novel group-shuffling random walk network is proposed to improve the training and testing processes by gallery-to-gallery affinities.

3. Proposed Method

We aim to optimize the architecture of the model to adaptively extract the potential salient feature of pedestrians. To this end, we proposed the *Salience-guided Cascaded Suppression Network* (SCSN). It introduces two new components: the feature aggregation modules (residual dual attention module and non-local fusion block) and the salient feature extraction unit. For convenience, in stage t, we denote the input feature map as X^t, the boosted feature map as Y^t, and the suppressed feature map for the input of $t+1$ stage as X^{t+1}. Our framework is illustrated in Fig. 2 .

3.1. Residual Dual Attention Module

The Residual Dual Attention Module (RDAM) consists of a Channel-wise Attention Module(CAM) and a Residual Spatial Attention Module (RSAM), in which the channel-wise attention module explores the correlation between channel features and the residual spatial attention module is responsible for exploring the semantically strong features within the spatial dimension.

Channel-wise Attention: The high-level convolutional feature in a trained CNN module is well-known to have remarkable localization ability for a semantic-related object. The channel-wise attention is introduced to enhance the representational ability for various pedestrians by explicitly modeling the interdependencies between the channel of convolutional features. To obtain the channel attention weight, we squeeze the spatial dimension of the input feature map by average pooling (to identify the extent of the object) and max pooling (to identify one discriminative part) simultaneously, generating two different 1D context descriptors: M^c_{avg} and M^c_{max}, which is similar to [44]. We then aggregate these descriptors via an attention mechanism [17] to obtain our channel attention map A_c. The detailed architecture of the attention agent is illustrated in Fig. 2(A). For a input, the channel attention vector is computed as:

$$A_c = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1 M^c_{avg}) + \mathbf{W}_2\delta(\mathbf{W}_1 M^c_{max})). \quad (1)$$

Herein, $\mathbf{W}_1$ and $\mathbf{W}_2$ are the parameters of FC layers and σ, δ denote the Sigmoid and ReLU function, respectively. The constructed channel-wise attention A_c is further applied to the original feature maps via channel-wise multiplication to enhance more informative channels and suppressing less useful ones. Then the obtained feature representations are further leveraged by the Residual Spatial Attention Module.

Residual Spatial Attention: The residual spatial attention is designed to guide the network to gather more necessary semantical information in the spatial dimension, which is complementary to the channel attention. To obtain the spatial attention map, we firstly aggregate the channel-wise

information of a feature map by two pooling operations: average pooling and max pooling, and generate two 2D maps: $M^s_{avg} \in \mathbb{R}^{H\times W}$ and $M^s_{max} \in \mathbb{R}^{H\times W}$. We then employ a convolution layer to aggregate M^s_{avg}, M^s_{max} and further obtain the spatial attention map $W_n \in \mathbb{R}^{H\times W}$ which encodes the locations to emphasize or suppress and n denotes the layer index of one stage. Consequently, inspired by [8], we allow the spatial attention information of previous blocks to propagate along with adjacent modules, named residual spatial aggregation, which enhances the consistency and robustness of spatial correlation estimation. Specifically, at each layer n in the same stage of backbone, A^s_n is the residual refined spatial map and A^s_{n-1} is the spatial map of the previous block. Then the update operation is defined as:

$$A^s_n = \sigma(A^s_{n-1} + \beta \cdot (W_n - A^s_{n-1})), \tag{2}$$

where β is a trainable variable initialized as 1, σ is the sigmoid activation function and we set $A^s_0 = 0$ for the first layer of each stage. Finally, A^s_n is applied to the input via an element-wise multiplication, as shown in Fig. 2(B).

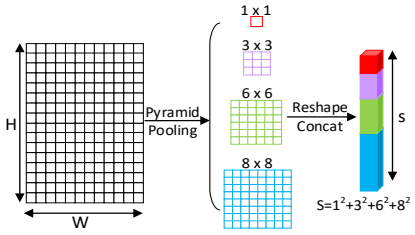

Figure 3. Illustration of the pyramid average pooling process.

3.2. Non-local Multistage Feature Fusion

Features fusion of different levels [20] has been demonstrated to be helpful to semantic segmentation, classification and detection. Common fusing operations are conducted in a pixel-wise, such as addition or concatenation, which has a limited performance gain because low-level features lack semantic information. To aggregate the features from different stages that are worth retaining for the final representation, we incorporate a non-local block to fuse multi-level features by leveraging long-range dependencies inspired by [54], called Non-local Multi-stage Feature Fusion (NMFF) block. Next, we elaborate on the detail of the NMFF block presented in Fig. 5. Specifically, we consider two types of source information for non-local fusion block: a high-level feature map $F_h \in \mathbb{R}^{C_h\times H_h\times W_h}$ and a low-level feature map $F_l \in \mathbb{R}^{C_l\times H_l\times W_l}$, where C, W and H denote the number of channel, width and height of features, respectively. Then, we employ three 1×1 convolutions ψ_q, ψ_v and ψ_k to transform F into compact embedding $F_q \in \mathbb{R}^{C'\times N_h}$, $F_v \in \mathbb{R}^{C'\times S}$ and $F_k \in \mathbb{R}^{S\times C'}$ as:

$$F_q = \psi_q(F_h), \quad F_k = \psi_k(F_l), \quad F_v = \psi_v(F_l), \tag{3}$$

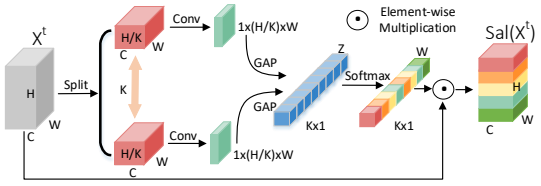

Figure 4. The detailed architecture of the *Salient Feature Extraction* (SFE) unit.

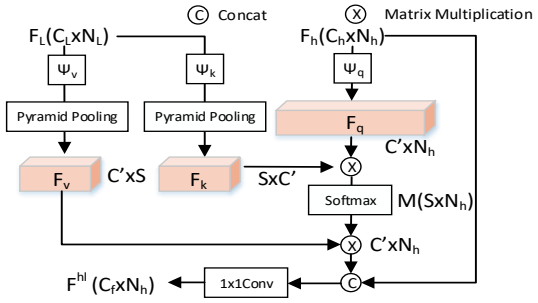

Figure 5. The detailed architecture of the *Non-local Multistage Feature Fusion* (NMFF) block.

where $N_h = W_h \times H_h$ and S represents pyramid average pooling pixels, as showed in Fig. 3. Consequently, we obtain the similarity matrix $M \in \mathbb{R}^{S\times N_h}$ by applying softmax on the matrix multiplication of F_k, F_q and then the fused output $F^{hl} \in \mathbb{R}^{C_f\times N_h}$ is computed by a matrix multiplication of M and F_v:

$$F^{hl} = M \times F_v = \mathcal{F}_{\text{softmax}}(F_k \times F_q) \times F_v. \tag{4}$$

Based on the above presentation, for the n stages employed for feature fusion, the final multi-stage fused feature F_f is obtained by

$$F_f = \phi(F^{hl_1}, F^{hl_2}, \ldots, F^{hl_n}), \tag{5}$$

where ϕ denotes a 1×1 convolution to reduce those features into a compact embedding.

3.3. Salient Feature Extraction Unit

As shown in Fig. 4, SFE unit behaves like a salience feature receptor that perceives which part-based features are discriminative. The unit can be decomposed into a salience descriptor and a salience selector.

Salience Descriptor: We divide the feature map into different stripes. Although we do not explicitly define the specific region features corresponding to each stripe of an object, this descriptor can be prone to guide each stripe to adaptively mine important information. As shown in Fig. 4, for a given input feature map $X^t \in \mathbb{R}^{C\times H\times W}$ at t stage, we uniformly split it into several part-based stripes and each of which has the size of $C\times(H/K)\times W$, where K denotes the number of stripes. Then, a convolutional layer followed by

batch normalization and ReLU is explored to capture fine-granularity information of each stripe, generating a compact feature descriptor with the shape of $1 \times (H/K) \times W$. Consequently, we apply a global average pooling operation on the feature descriptor to get the feature vector $\mathbf{z} \in \mathbb{R}^{k \times 1}$. Obviously, the larger the number of part-based stripes is, the finer the feature descriptor is.

Salience Selector: After obtaining the feature vector $\mathbf{z}$, we employ the salience selector that consists of a softmax activation and an element-wise multiplication $\odot$, which is analogous to an attention mechanism. Then we can get the salience-sensitive weights $W = (w_1, ..., w_k)^T$ and salience local feature $\mathcal{S}al(X^t)$:

$$\mathcal{S}al(X^t) = W \odot X^t, \tag{6}$$

$$w_i = \frac{\exp(z_i)}{\sum_1^k \exp(z_j)}, \quad i \in [1, k]. \tag{7}$$

$\mathcal{S}al(X^t)$ is highlighted in stage t while will be suppressed in stage $t+1$.

3.4. Salience-Guided Cascaded Suppression Network

Multi-Stage Suppression: Our proposed SCSN employs ResNet50 as the backbone. Notice that we modified the downsample strides of $Stage3$ and $Stage4$ to 1 to preserve more spatial information. After getting the basic feature from the backbone, we extract potential salience feature stage-by-stage. Specifically, for stage t, we first extract salient feature $\mathcal{S}al(X^t)$ of this stage by SFE unit and then the $\mathcal{S}al(X^t)$ will be integrated with the base input feature X^t as follow:

$$Y^t = X^t + \mathcal{S}al(X^t), \tag{8}$$

where Y^t denotes the salience boosted feature. The promotion of $\mathcal{S}al(X^t)$ alleviates the dilution of detail information due to the global average pooling and the summation integration method also avoids the dimensional inefficiency caused by concatenation. Consequently, to mine other potential salient features, we apply a salience mask on the output of stage t to suppress $\mathcal{S}al(X^t)$ and obtain the input X^{t+1} of stage $t+1$:

$$X^{t+1} = X^t \cdot \mathcal{B}(X^t), \tag{9}$$

where $\mathcal{B}$ is a binary mask which takes values of the most salient $\mathcal{S}al(X^t)$ to 0 and others to 1. The suppression operation relieves the coverage effect of $\mathcal{S}al(X^t)$ on other features and makes potential information stand out. Therefore, the network can further discover more potential features. The detailed pipeline of SCSN is shown in Fig. 2. We consider backbone's last convolution block followed by a SFE unit as the global ($t = 1$) stage. The feature extracted in the global context is the most salient one among all stripe-based local features. The following stages then continue to mine salience features in the same manner. Fig. 6 shows an intuitive salient feature visualization of 4 hard samples with similar appearances. To avoid information redundancy, we first apply global pooling on the boosted output Y^t to generate a 2048-dimensional feature vector and then use the FC layer to reduce the vector dimension. Note that, in our experiments, the global stage uses the average pooling to get the feature vector while the following stages use the max pooling because salience suppression operation causes instability of the mean of the features.

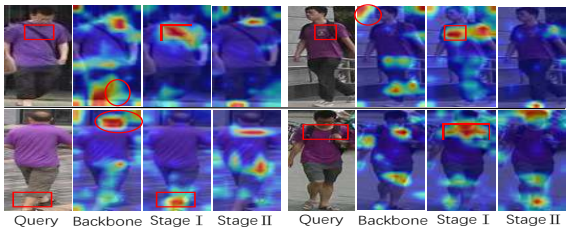

Figure 6. The feature visualization for 4 hard samples with similar appearances. Warmer color denotes higher value. We can observe that the backbone features are not accurate enough such as the interference of background, as shown in the red circles. However, the stage I features extracted by SFE unit focus on such salient features, as noted in red boxes. And, after the salience feature suppression, stage II discovers some potential features which are also important but masked by salient features of stage I.

Comparison with Feature Erasing: Inspired by the cognitive process of humans, 'deliberately' suppressing the areas we are not interested in is to better focus on our attention [21]. The similar idea of feature erasing has also been used for weakly-supervised object localization and semantic segmentation [15, 46]. However, the purpose and implementation of these methods are different from ours. Specifically, by erasing the object-related region that the network already captured, these methods encourage CNN to discover more semantic areas related to object for the integral object attention. On the contrary, the proposed SFE unit aims to extract the most salience feature that can distinguish one instance to others, looking for a salience subset in the basis feature space instead of searching for a more complete feature space as [15, 46] did.

Moreover, our approach is more labor-saving than other methods that rely on auxiliary information, such as attribute learning [24] and pose/human parsing [19, 25, 29]. Specifically, with the large labeled datasets, attribute learning methods force the network to pay attention to local contexts by the supervision of specific labels. But the large number of attribute-labels cannot always be taken for granted. Meanwhile, pose/human parsing related methods utilize the prior human body knowledge that comes from the pose estimation and semantic parsing algorithms. They are prone

to be misguided by the noise of prior because semantic segmentation and pose estimation themselves are challenging tasks in complex scenarios.

Loss Function: Identification loss gets the ID prediction logits of images, which is similar to the classification loss, defined as:

$$L_{id} = \sum_{i=1}^{N} -q_i \log(p_i) \begin{cases} q_i = \varepsilon/N & y \neq i, \\ q_i = 1 - \varepsilon\frac{N-1}{N} & y = i, \end{cases} \tag{10}$$

where y and p_i denote the ground-truth ID label and the predicted logit of class i, respectively. N represents the number of classes and q_i is the smoothed label which is proposed in [37] and $\varepsilon = 0.1$ is used to smooth the label.

In addition, considering the property of Re-ID, that is, finding the most similar series of people from the gallery, we introduce the idea of metric learning, which enables the network to find features that are useful for similarity metrics[13]. Therefore, we employ the triplet loss to enhance the final ranking performance, which is defined as:

$$L_{tp} = \sum\nolimits_{N} [d_p - d_n]_+ , \tag{11}$$

where d_p is the feature distance of the same identity and d_n denotes the distance of different identities. N is the batch-size of triplet samples and $[\cdot]_+$ means $\max(\cdot, 0)$. Triplet loss aims to ensure that the distance between positive sample pairs is less than the distance between negative sample pairs. The final loss of our model can be written as:

$$L = L_{id} + L_{tp}. \tag{12}$$

Finally, we jointly train the end-to-end multi-staged suppression network with intermediate supervision loss.

4. Experiments

4.1. Implementation

Experimental details: We resize all images into the resolution of 384×128 and set the number of stripes $K = 8$. Our backbone is ResNet50 pre-trained on ImageNet. For data augmentation, we deploy random horizontal flipping and random erasing in the training dataset, which is the same as [26]. For each iteration, mini-batch sampled to the triplet loss consists of $B = P \cdot I$ images, including randomly selected P identities and randomly sampled I images for each identity. Here we take $P = 48$ and $I = 4$. For NMFF block, we take $c_f^{34} = 2048$ and $c^{'} = 256$ for Market1501, and set $c_f^{14} = c_f^{24} = c_f^{34} = 512$, $c^{'} = 64$ for the rest three datasets. We employ Adam as the optimizer with the weight decay factor of 0.0005. In addition, a warmup strategy is applied to make the training gradient smooth. In practice, we first spend 20 epochs to linearly increase the learning rate from 1.25×10^{-4} to 2.5×10^{-3} (20 times) and then employ a cosine annealing strategy to decrease the learning rate from 2.5×10^{-3} to 0, using another 200 epochs.

Method	Backbone	mAP	rank 1
SCSN(4 stages)	ResNet50	88.30	92.40
SCSN(3 stages)	ResNet50	**88.50**	**95.70**
*ABDNet [7](ICCV19)	ResNet50	88.28	95.60
†Pyramid[49](CVPR19)	ResNet101	88.20	**95.70**
DCDS[1](ICCV19)	ResNet101	85.80	94.81
* †MHN(PCB) [5](ICCV19)	ResNet50	85.00	95.10
†MGN [24] (ACM MM 18)	ResNet50	86.90	95.70
BFE [10](ICCV19)	ResNet50	86.20	95.30
*†CASN(PCB)[51](CVPR19)	ResNet50	82.80	94.40
*†AANet[38](CVPR19)	ResNet152	83.41	93.93
*IANet[16](CVPR19)	ResNet50	83.10	94.40
*†VPM[34](CVPR19)	ResNet50	80.80	93.00
§PSE+ECN[29](CVPR18)	ResNet50	80.50	90.40
†PCB+RPP[36](ECCV18)	ResNet50	81.60	93.80
†PCB[36](ECCV18)	ResNet50	77.40	92.30
*DuATM [31](CVPR18)	DenseNet121	76.60	91.40
§Pose-transfer[25](CVPR18)	DenseNet169	56.90	78.50
§SPReID[19](CVPR18)	ResNet152	83.36	93.68
Tricks[26](CVPRW19)	SEResNet101	87.30	94.60
*Mancs[40](ECCV18)	ResNet50	82.30	93.10
PAN[53](TCSVT18)	ResNet50	63.40	82.80
SVDNet[35](CVPR17)	ResNet50	62.10	82.30

∗ Attention related, † Stripes related, § Pose or human parsing related.

Table 1. Comparison with state-of-the-art person Re-ID methods on the Market1501 dataset.

Protocols: To compare the performance of the proposed method with the existing advanced Re-ID methods, we adopt the Cumulative Matching Characteristics (CMC) at rank1 and mean Average Precision (mAP) as the evaluation metrics for each query image on 4 datasets. It is worth noting that re-ranking is **not** used for simplicity.

4.2. Datasets

Market1501: It [50] consists of 32668 images of 1501 individuals in total, among which 12936 images of 751 persons are used as the training set. And 19732 images of 750 people are separated into the testing set with 3368 query images as well as 16364 gallery images. Besides, this dataset is shot by six different cameras with bounding-boxes annotated by the Deformable Part Model (DPM) detector [11].

DukeMTMC-ReID: It is a subset of the DukeMTMC [52] and is also a popular dataset for human Re-ID, including 36411 images of 1812 identities from eight high-resolution cameras. Specifically, the training set contains 16522 images of 702 identities that are randomly selected from the overall images, and the testing set comprises the other 2228 query images and 17661 gallery ones.

CUHK03: This dataset [22] comprehends 14097 images of 1467 persons from ten cameras and is divided into the training set of 767 individuals and the testing set of 700 individuals. Besides, it provides two types of annotations,

Method	Backbone	mAP	rank 1
SCSN(4 stages)	ResNet50	**79.00**	**91.00**
SCSN(3 stages)	ResNet50	**79.00**	90.10
†Pyramid [49](CVPR19)	ResNet101	**79.00**	89.00
*ABDNet [7](ICCV19)	ResNet50	78.60	89.00
†MGN [24] (ACM MM 18)	ResNet50	78.40	88.70
*†MHN(PCB)[5](ICCV19)	ResNet50	77.20	89.10
BFE [10](ICCV19)	ResNet50	75.90	88.90
* †CASN(PCB)[51](CVPR19)	ResNet50	73.70	87.70
DCDS[1](ICCV19)	ResNet101	75.50	87.50
* †AANet[38] (CVPR19)	ResNet152	74.29	87.65
§PSE+ECN[29](CVPR18)	ResNet50	75.70	84.50
*IANet[16](CVPR19)	ResNet50	73.40	83.10
*VPM[34](CVPR19)	ResNet50	72.60	83.60
*DuATM [31](CVPR18)	DenseNet121	64.60	81.80
†PCB+RPP[36](ECCV18)	ResNet50	69.20	83.30
§SPReID[19](CVPR18)	ResNet152	73.34	85.95
§Pose-transfer[25](CVPR18)	DenseNet169	56.90	78.50
Tricks[26](CVPRW19)	SEResNet101	78.00	87.50
*Mancs[40](ECCV18)	ResNet50	82.30	93.10
SVDNet[35](CVPR17)	ResNet50	56.80	76.70
PAN[53](TCSVT18)	ResNet50	51.51	71.59

∗ Attention related, † Stripes related, § Pose or human parsing related

Table 2. Comparison with state-of-the-art person Re-ID methods on the DukeMTMC-ReID dataset.

i.e., manually labeled pedestrian boxes and DPM detected boxes. More concretely, the manually labeled part consists of 7368 training images, 1400 query images and 5328 gallery images; the DPM detected part includes 7365 training images, 1400 query images and 5332 gallery images.

MSMT17: It [42] is a new public person Re-ID dataset. There are 126441 images of 4101 identities captured by a 15-camera network, including 12 outdoor and 3 indoor, in this largest dataset. And all the boxes are annotated by Faster RCNN [28]. Therefore, MSMT17 is more challenging than other public person Re-ID datasets, due to its massive scale, more complex and dynamic scenes.

4.3. Comparison with State-of-the-Art Methods

We compare the proposed SCSN with current state-of-the-art methods on four datasets, including Market1501, DukeMTMC-ReID, CUHK03 and MSMT17 to demonstrate the robustness and the superior performance over other advanced methods. Results are given as following.

Market1501: Table 1 shows the results of Market1501. These methods are divided into two groups: the top of Table 1 are methods that integrate local features and global features, called the global-local group and the bottom are methods merely employing global features, called the global group. Our SCSN obtains the best mAP performance and the same rank1 result as Pyramid [49]. However, it is worth pointing out that Pyramid benefits from a larger backbone and a more complex pyramidal feature set (including 21 stripe features). Tricks is the representative of the global group. With the strong backbone SEResNet101,

Method	Backbone	Labeled		Detected	
		mAP	rank1	mAP	rank1
SCSN(4 stages)	ResNet50	**84.00**	**86.80**	**81.00**	**84.70**
SCSN(3 stages)	ResNet50	83.30	86.30	80.20	84.10
†Pyramid[49](CVPR19)	ResNet101	76.90	78.90	74.80	78.90
BFE[10](ICCV19)	ResNet50	76.70	79.40	73.50	76.40
*†MHN(PCB)[5](ICCV19)	ResNet50	72.40	77.20	65.40	71.70
†MGN[24] (ACM MM 18)	ResNet50	67.40	68.00	66.00	68.00
†PCB+RPP[36](ECCV18)	ResNet50	-	-	57.50	63.70
* †CASN(PCB)[51](CVPR19)	ResNet50	68.00	73.70	64.40	71.50
Tricks[26](CVPRW19)	SEResNet101	70.40	72.00	68.00	69.60
*Mancs[40](ECCV18)	ResNet50	63.90	69.00	60.50	65.50
SVDNet[35](CVPR17)	ResNet50	37.80	40.90	37.30	41.50
PAN[53](TCSVT18)	ResNet50	35.00	36.90	34.00	36.30

∗ Attention related, † Stripes related

Table 3. Comparison with state-of-the-art person Re-ID methods on the CUHK03 dataset with the 767/700 split.

Tricks achieves comparative results. However, the inherent defect of the global feature, that is, the dilution of the local significant information, limits further improvement of such methods. Finally, our proposed cascaded feature suppression strategy achieves further improvements, benefitting from the employ of potential salient features explored by SFE units.

DukeMTMC-ReID: The results of this dataset are shown in Table 2. Similar to Market1501 dataset, the proposed SCSN also achieves the best results on rank1 and exceeds Pyramid [49] / ABDNet [7] by 2.0%. Besides, we achieve the same mAP performance as Pyramid [49] while uses a lightweight backbone.

CUHK03: It is a more challenging dataset, compared with Market1501 and DukeMTMC-ReID. This is reflected in **1)** CUHK03 has fewer samples and the viewpoint variations and the occlusion problems are serious; **2)** the annotation of bounding boxes marked by object detection algorithm has location offsets. While, as reported in Table 3, SCSN substantially exceeds Pyramid [49] / BFE [10] by 7.1%/7.3% in mAP and 7.9%/7.4% in rank1 metric on the labeled dataset and exceeds the performances of Pyramid [49] / BFE [10] by 6.2%/7.5% in mAP and 5.8%/8.3% in rank1 on the detected dataset, achieving the most outstanding results. The experimental results clearly demonstrate that under the condition of limited training samples, mining potential features and integrating these complementary features enjoys great advantages. Specifically, with limited samples, the attention mechanism is prone to partially learn the most significant features, but the learned biased feature maybe not robust and essential because it will change and even be lost during testing. Meanwhile, indiscriminate stacking redundant local features, such as Pyramid [49], can only bring limited improvements.

MSMT17: Table 4 shows the results of MSMT17 which is the latest and largest Re-ID dataset and the amount of methods that report on this dataset is less because it is recently publicly-available. SCSN achieves the best performances on rank1 and rank5 metrics.

Method	Backbone	mAP	rank1	rank5
SCSN(4 stages)	ResNet50	58.50	**83.80**	**91.50**
SCSN(3 stages)	ResNet50	58.00	83.00	91.20
ABDNet[7](ICCV19)	ResNet50	**60.80**	82.30	90.60
BFE[10](ICCV19)	ResNet50	51.50	78.80	89.10
IANet[16](CVPR19)	ResNet50	46.80	75.50	85.50
GLAD[43](ACM MM 17)	ResNet50	34.00	61.40	76.80
PDC[32](ICCV17)	GoogLeNet	29.70	58.00	73.60
ResNet50[13](CVPR16)	Baseline	33.90	63.20	-

Table 4. Comparison with state-of-the-art person ReID methods on the MSMT17 dataset.

4.4. Ablation Studies

To demonstrate the effects of feature aggregation and feature suppression blocks in SCSN, we incrementally evaluate each module on DukeMTMC-ReID. We take ResNet50 with the global branch as the baseline, employing the ID loss and Triplet loss. Nine variants are then conducted based on the baseline: **a)** baseline + CAM; **b)** baseline + SAM (without residual update); **c)** baseline+ RSAM ; **d)** baseline+ CAM +SAM ; **e)** baseline + RDAM (CAM + RSAM); **f)** B&A + SFE; **g)** B&A + SFE$*$1; **h)** SCSN (B&A + SFE$*$2) and **i)** SCSN (B&A + SFE$*$3), where B&A represents the backbone equipped with the residual dual attention attention module. Table 5 presents the ablation study results, from which several observations could be drawn:

1) All these three attention modules: CAM, Spatial Attention Module(SAM), and RSAM improve the baseline and the RDAM that combines the CAM and RSAM mechanisms can further improve performances, which demonstrates their complementary property and the feature extraction ability. Specifically, the RSAM achieves better performance than SAM. Because the spatial attention module trained in a weakly-supervised manner, without residual connection, cannot obtain accurate position attention maps [7, 15]. There are two reasons: 1) the weakly-supervised training manner, without a powerful supervisory, result that gradient of SAM might be vanishing in the backpropagation process, especially for modules in shallow layers; 2) compared with higher blocks that have rich semantic features, such as the human body, shallow layers focus more on the low-level feature, such as color and texture, which pose an obstacle to get accurate context attention maps.

2) By combining the proposed cascaded feature suppression strategy and attention mechanisms, performance is further boosted. Specifically, in our experiments, 3 or 4 cascaded stages achieve the best performance. Too many cascaded stages can lead to inferior results. We argue that for a pedestrian, its distinctive features are limited so excessive suppressions force the network to learn some non-robust features. Besides, since the features of different stages are contacted during the testing, too many stages can cause the feature vectors redundant, diluting the salient features and increasing the computation amount.

Method	Backbone	mAP	rank 1
Baseline	ResNet50	71.50	86.10
Baseline + CAM	ResNet50	73.80	87.40
Baseline + SAM(without residual)	ResNet50	73.50	86.60
Baseline + RSAM	ResNet50	74.10	87.30
Baseline + CAM+SAM	ResNet50	75.80	87.80
Baseline + RDAM(CAM+RSAM)	ResNet50	76.20	88.90
B&A + SFE*1	ResNet50	78.30	89.50
B&A + SFE*2	ResNet50	78.80	89.80
B&A + SFE*3	ResNet50	78.50	90.20
SCSN(B&A+SFE*2+NMFF)	ResNet50	79.00	90.10
SCSN(B&A+SFE*3+NMFF)	ResNet50	**79.00**	**91.00**

Table 5. Ablation study of SCSN on DukeMTMC-ReID dataset. Here NMFF denotes the non-local multistage feature fusion block with all four stages in ResNet50.

Method	B	B+N(14)	B+N(24)	B+N(34)	B+N(24,34)	B+N(14,24,34)
mAP	78.50	78.60	78.10	77.80	78.02	**79.00**
rank1	89.50	89.90	89.30	89.50	89.54	**90.10**

Table 6. The effect of different feature fusion strategies, where N denotes the NMFF block and B denotes the baseline without feature fusion. The number of combinations represent the fused feature of different stages of ResNet50. As illustrated, merely merging advanced features, such as (24) and (34), does not bring too much gain because these advanced features have more similar distributions. On the other hand, the fusion of shallow features can significantly improve the accuracy of the network.

3) The feature aggregation modules and the suppression mechanism are complementary. Specifically, without the aggregation module improving the context-awareness of the backbone, the suppression mechanism may focus on the interference information in the background. In contrast, without the suppression modules, the attention and aggregation mechanisms tend to only learn the most salient features that can distinguish each identity and ignore other important information. More detailed differences in feature fusion strategies of NMFF are shown in Table 6.

5. Conclusion

In this paper, we propose a novel method called Salience-guided Cascaded Suppression Network (SCSN) for person re-identification which solves the problems of how to extract discriminative features and how to integrate these features. The suppression strategy can be considered as a salient dropout scheme that enables the network to adaptively mine potential significant information on different important levels. Extensive experiments demonstrate that our method achieves state-of-the-art results on 4 popular person Re-ID benchmarks, and it is worth noting that our proposed method makes over 7% improvement on CUHK03 datasets. In the future, it is meaningful to investigate a more effective feature extraction method.

Acknowledgement This work is supported in part by the National Natural Science Foundation of China under Grant 61972188, the Science and Technology Planning Project of Shenzhen (No. JCYJ20180503182133411).

References

[1] Leulseged Tesfaye Alemu, Marcello Pelillo, and Mubarak Shah. Deep constrained dominant sets for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9855–9864, 2019.

[2] Igor Barros Barbosa, Marco Cristani, Barbara Caputo, Aleksander Rognhaugen, and Theoharis Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *Computer Vision and Image Understanding*, 167:50–62, 2018.

[3] Binghui Chen and Weihong Deng. Energy confused adversarial metric learning for zero-shot image retrieval and clustering. *arXiv preprint arXiv:1901.07169*, 2019.

[4] Binghui Chen and Weihong Deng. Hybrid-attention based decoupled metric learning for zero-shot image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2750–2759, 2019.

[5] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. *arXiv preprint arXiv:1908.05819*, 2019.

[6] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2018.

[7] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8351–8361, 2019.

[8] Xuesong Chen, Haihua Lu, Kaili Cheng, Yanbo Ma, Qiuhao Zhou, and Yong Zhao. Sequentially refined spatial and channel-wise feature aggregation in encoder-decoder network for single image dehazing. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2776–2780, 2019.

[9] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1335–1344, 2016.

[10] Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3691–3701, 2019.

[11] Pedro F Felzenszwalb, David A McAllester, Deva Ramanan, et al. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[12] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 262–275, 2008.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[15] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 549–559, 2018.

[16] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9317–9326, 2019.

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[18] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5098–5107, 2018.

[19] Mahdi M. Kalayeh, Emrah Basaran, Muhittin G?kmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, 2006.

[21] Fei Fei Li, Rufin VanRullen, Christof Koch, and Pietro Perona. Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences (PNAS)*, pages 9596–9601, 2002.

[22] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 152–159, 2014.

[23] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. *arXiv preprint arXiv:1705.04724*, 2017.

[24] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017.

[25] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4099–4108, 2018.

[26] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 0–0, 2019.

[27] Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary. Person re-identification by support vector ranking. In *The British Machine Vision Conference (BMVC)*, volume 2, page 6, 2010.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.

[29] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 420–429, 2018.

[30] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2265–2274, 2018.

[31] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5363–5372, 2018.

[32] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3960–3969, 2017.

[33] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3960–3969, 2017.

[34] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 393–402, 2019.

[35] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3800–3808, 2017.

[36] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[38] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7134–7143, 2019.

[39] Evgeniya Ustinova, Yaroslav Ganin, and Victor Lempitsky. Multi-region bilinear convolutional neural networks for person re-identification. In *Proceedings of 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.

[40] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 365–381, 2018.

[41] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of ACM Multimedia Conference on Multimedia Conference (ACM MM)*, pages 274–282, 2018.

[42] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 79–88, 2018.

[43] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *Proceedings of the 25th ACM international conference on Multimedia (ACM MM)*, pages 420–428, 2017.

[44] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[45] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing (TIP)*, 28(6):2860–2871, 2019.

[46] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1325–1334, 2018.

[47] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3219–3228, 2017.

[48] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3586–3593, 2013.

[49] Feng Zheng, Deng Cheng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Ji Rongrong. Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[50] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International*

Conference on Computer Vision (ICCV), pages 1116–1124, 2015.

[51] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J Radke. Re-identification with consistent attentive siamese networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5735–5744, 2019.

[52] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3754–3762, 2017.

[53] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2018.

[54] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 593–602, 2019.