A CAM-Guided Parameter-Free Attention Network for Person Re-Identification

A CAM-guided Parameter-free Attention Network for Person Re-Identification

Wanlu Li, Yunzhou Zhang, *Member, IEEE,* Sonya Coleman and Weidong Shi

***Abstract*—In this paper, we propose a parameter-free attention mechanism based on class activation mapping (CAM) which is novel compared with most of the existing works that train attention without a supervision signal. Our attention is composed of spatial attention and channel attention in the standard way, which indicates that "where" and "what" is more meaningful, respectively. For Spatial Attention, we use class activation mapping as a supervision signal to guide the generation of it directly in space. Thus our approach to spatial attention can pay more attention to the informative pedestrian parts of the scene and reduce background interference. For Channel Attention, the importance of each channel is obtained by the similarity between the aforementioned spatial attention and the feature map of each channel. In this manner, our channel attention is indirectly guided by CAM. In addition, our attention is parameter-free, which reduces the risk of overfitting. Finally, we conduct extensive evaluations on three popular benchmark datasets including Market1501, DukeMTMC-reID, and MSMT17, demonstrating the effectiveness of our approach on discriminative person representations.**

***Index Terms*—Person re-identification, attention mechanism, Class Activation Mapping, parameter-free**

I. Introduction

THE intent of person re-identification (Re-ID) is to match certain people from the images taken by non-overlapping surveillance cameras. Person Re-ID has sparked huge interest from academia and industry as it plays an important role in video surveillance. With the development of deep learning, person Re-ID models based on deep Convolution Neural Networks have made remarkable progress. However, person Re-ID still faces enormous challenges owing to occlusion, cluttered background, the similarity of pedestrian appearance. To solve the aforementioned problem, recent research [1]–[5] has applied the attention mechanism to person Re-ID, because of its advantage of emphasizing discriminative features and suppressing the interference of irrelevant features. In general, the attention mechanism includes channel attention and spatial attention.

Spatial attention reflects the importance of pixels, which helps the network to focus on people regions and suppress the influence of cluttered background. Chen *et al*. [4] propose ABD-Net, which applies the non-local form of self-attention to person Re-ID, and achieves position awareness by computing a pixel affinity matrix. Xia *et al*. [6] replace the affinity matrix with a covariance matrix to obtain second-order statistics used for modelling long-range relationships. However, there is no strong supervision signal to guide the generation of the spatial mask which limits the performance of the model when extracting discriminative features. However, this can be overcome by supervised attention methods. Chen *et al*. [7] designed a critic inside the attention module to judge whether the attention leads to a correct classification and ameliorates the representation. Zhu *et al*. [8] apply spatial attention to a global average pooling (GAP) layer to obtain a feature vector, and then compare the feature vector with the pedestrian label to realize the supervision. Nevertheless, these methods do not directly supervise the attention map in space and may suffer if the spatial information is missing. Hence, we propose a method which directly supervises spatial attention based on Class Activation Mapping (CAM) [9]. CAM is a common visualization tool that has been used to display discriminative regions of an input image. We directly extract the feature map corresponding to the pedestrian identity to obtain CAM. By calculating the Mean-Square Error (MSE) loss of CAM and spatial attention, the label-related spatial attention is obtained in the end-to-end training phase.

The role of channel attention is to assign weights to different channels. Hu *et al*. [10] compress spatial information and obtain channel descriptors by squeeze and excitation, respectively. To control the complexity of the model, the excitation operation contains dimensionality reduction which has an impact on the attention prediction. Therefore, Wang *et al*. [11] proposed ECA-Net to capture local cross-channel interactions by considering each channel and its k-nearest neighbors. It reduces the computational complexity while avoiding dimensionality reduction. However, it calculates the local dependencies and cannot obtain the relationship between all channels and classification categories. In person Re-ID research, ABD-Net [4] computes channel attention by designing a channel affinity matrix with reference to non-local spatial attention. Nevertheless, these approaches lack a powerful supervised signal that would establish the connection between channels and pedestrian labels. Additionally, channel attention often needs to be inserted into the network many times which increases computation. In order to associate the channel with pedestrian identities and avoid multiple embedding, we propose a simple yet effective approach for channel attention. We measure the effectiveness of the region that the channel focuses on by the similarity between the feature map of each channel and the CAM. Specifically, we calculate the cosine similarity between the CAM-guided spatial attention and feature maps, and map it to the Gaussian space to obtain the nonlinear channel weights. In this manner, the channels

The authors are with Institute of Image Recognition and Machine Intelligence, College of Information Science and Engineering, Northeastern University, Shenyang 110819, China (e-mail:1900825@stu.neu.edu.cn; zhangyunzhou@mail.neu.edu.cn; shiweidong1003@gmail.com)

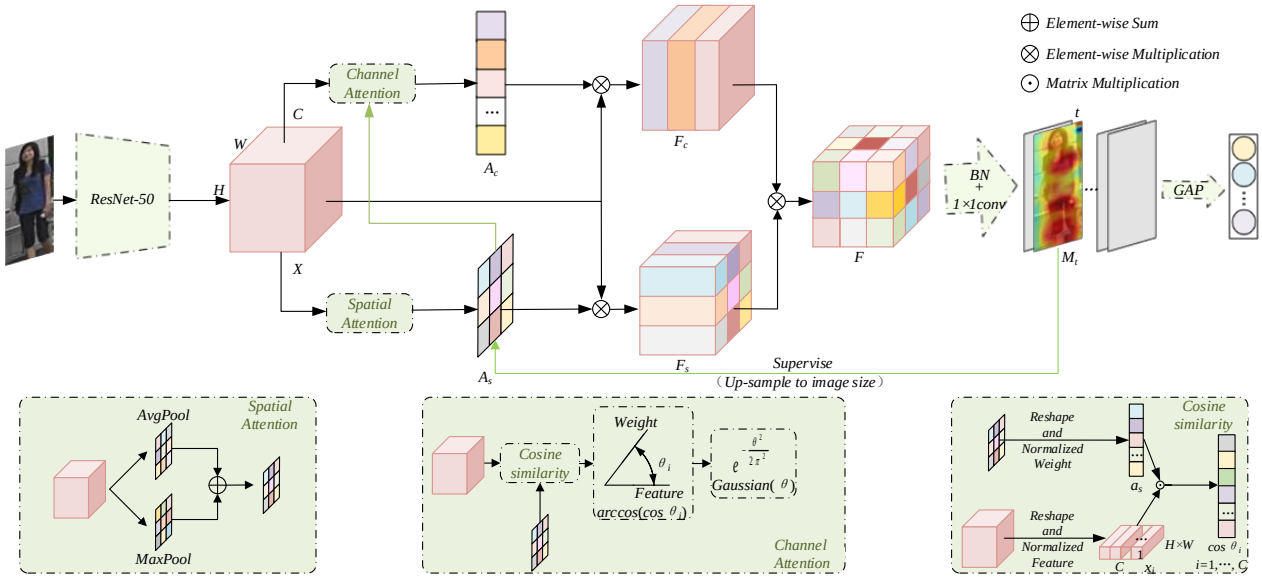

Fig. 1. The pipeline of the proposed framework. The top part shows the overall architecture of the network, and the bottom part shows the specific implementation details of the attention module

with high similarity to the spatial attention corresponding to the pedestrian identity are assigned a larger weight. In this way, we obtain the label-related channel attention based on CAM.

Moreover, unlike the above mentioned studies that add additional trainable parameters, our work increases the performance of the model without extra parameters. This reduces the risk of model over-fitting. In contrast to the parameter-free work [12] that only proposes a spatial attention and needs to be embedded in the first three stages of the network, we propose both simpler spatial with pooling operations only and channel attention that do not require multiple embedded instances in the model. In summary, the contribution of the letter is as follows:

- We design the CAM as a supervisory signal that directly guides spatial attention to pedestrian positioning in space without the need for additional parameters.
- We propose a parameter-free channel attention to establish the relationship between channel and label and to improve nonlinear feature extraction capability.
- Our attention model is plug-and-play, and we validate the efficacy of our method by conducting extensive experiments on three well-known person Re-ID datasets.

II. The Proposed Method

A. Baseline

We adapt the *stage1-stage4* of ResNet50 [13] as the backbone for feature extraction. We remove the down-sampling operation of res_conv4a [14] to obtain richer features. In order to obtain the CAM more readily, we adjust the classifier in ResNet50. Firstly, the original full connected layer is removed, and then a 1×1 convolution layer with the number of output channels as the number of pedestrian identities is added before the global average pooling (GAP). In this way, the CAM M_t is directly obtained by selecting the feature map corresponding to the current image identity t during the forward propagation. In addition, we add a batch normalization (BN) layer [15] between the feature extraction and the classification layers to achieve better performance than the current state-of-the-art approaches.

B. Spatial Attention

Spatial attention encodes where to emphasize or suppress information by assigning weights to each pixel based on feature importance. As shown in the bottom left of Fig. 1, we aggregate channel information from feature maps by applying both average-pooling and max-pooling operations along the channel axis. In contrast to [16], instead of utilizing a convolutional layer to integrate the two features obtained by pooling, we directly add them to obtain the spatial attention $A_s \in \mathbb{R}^{1\times H\times W}$. With this approach, our proposed model does not contain any additional parameters. Therefore, the spatial attention is computed as

$$A_s(X) = N[N(Avg(X)) + N(Max(X))] \tag{1}$$

where $Avg()$ is the average pooling, $Max()$ is the max pooling, and $N()$ denotes the normalized operation, calculated as:

$$N(f) = \frac{x - min(f)}{max(f) - min(f)} \tag{2}$$

Previous approaches to spatial attention lack effective supervision, which limits their calibration ability in space, hence we design a spatial attention supervision approach based on CAM. CAM typically has a higher response in a pedestrian region of an image, and has a very low response in the background area [17]. Its spatially weighted property allows it to serve as a strong supervisory signal for spatial attention. We judge whether spatial attention is effective by calculating its similarity to CAM. If there is a significant difference between

them, then the attention map is not effectively focused on the discriminative region of the pedestrian image.

C. Channel Attention

The feature maps of each channel extract information from different spatial regions, and the channel attention reflects the degree of this information contribution to classification. In person Re-ID, the discriminative features of the image are concentrated around the pedestrian region and the cluttered background information should be ignored. Thus the channel that extracts information from the pedestrian region should have a larger weight and the channel that extracts information from the background region should be assigned a smaller weight. Since our spatial attention is supervised by CAM, to focus more on the pedestrian region, the channel attention can be obtained by calculating the correlation between the feature maps of each channel and the spatial attention. We describe the detailed operation below.

First, we calculate the cosine similarity of the feature map X to the spatial attention A_s, as shown in the bottom right of Fig.1. Here we need to convert the feature map $X \in \mathbb{R}^{C\times H\times W}$ into feature vectors $x_i \in \mathbb{R}^{(H\times W)\times 1}, i \in 1, 2, \cdots, C$ and the spatial attention $A_s \in \mathbb{R}^{1\times H\times W}$ into vectors $a_s \in \mathbb{R}^{1\times (H\times W)}$. The cosine similarity is calculated as

$$cos\theta_i = \frac{a_s \cdot x_i}{||a_s||||x_i||} i \in 1, 2, \cdots, C \tag{3}$$

Then we take the inverse trigonometric cosine function to obtain θ_i, and finally it is substituted into the Gaussian function with $\mu = 0, \sigma = \pi$ to acquire the channel attention $A_c \in \mathbb{R}^{C\times 1\times 1}$, which can be formulated as

$$A_c(X) = Gaussian(\theta) = e^{-\frac{\theta^2}{2\pi^2}} \tag{4}$$

In this design, the more similar the channel feature map is to the spatial attention, the closer the angle θ_i between the two vectors is to 0°, and distributed near the vertical axis of the Gaussian function. Also, as the output of the Gaussian function increases, so too does the response of the corresponding channel, and vice versa.

Our channel attention improves the ability of our proposed model to extract nonlinear features. As CAM is the supervised signal of spatial attention, our channel attention is supervised by CAM in the process of calculating similarity.

D. Loss

To guide attention generation, we define the attention loss using mean square error (MSE) between the spatial attention map and CAM. The formula is as follows

$$L_a = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}||M_t(h,w) - A_s(h,w)||^2 \tag{5}$$

where the sizes of A_s and M_t are uniformly up-sampled to the image size, *i.e.*, $H \times W$, and the overall objective loss function is formulated as

$$L = L_{id} + L_{tri} + \lambda L_a \tag{6}$$

where L_{id} is the identity loss [18], L_{tri} is triplet loss [19], L_a is the attention loss and λ is hyper-parameter for balancing the three losses. We finally optimize the network model by minimizing L.

III. Experiments

A. Datasets and Evaluation Metrics

We verify the effectiveness of the proposed approach on three large-scale public datasets, Market-1501 [20], DukeMTMC-ReID [21], and MSMT17 [22], by comparing with the baseline method and the state-of-the-art method. The Market1501 dataset contains 12936/19732 images of 751/750 persons for training/testing captured by 6 cameras. DukeMTMC-ReID is taken by 8 high-resolution cameras, and the training set contains 16522 images of 702 identities, the testing set includes 2228 query images and 17661 gallery images of another 702 identities. MSMT17 consists of 32621 images of 1041 persons as the training set and 93820 images of 3060 people as the testing set collected across 15 cameras, including 12 outdoor and 3 indoor. This dataset is more challenging because of its massive scale, complex scenarios, and large variation of illumination. We adopt the Cumulative Matching Characteristics (CMC) for rank-1 and mean Average Precision (mAP) as the evaluation protocol by convention.

B. Implementation Details

The proposed approach is implemented using a Pytorch framework with only one Nvidia TITAN GPU. The input images are resized to 384×128. During the training phase, we use the Adam optimizer and set the weight decay to 0.0005. The batch size is set to 32, with 4 identities in each mini-batch and 8 images for each identity. We fine-tune the classifier parameters using a total of 60 epochs. Among them, the first 10 epochs use warmup learning rate, which gradually increases from 3.5×10^{-6} to 3.5×10^{-4}. After 30 and 50 epochs, the learning rate is dropped 0.1×, *i.e.*, 3.5×10^{-5} and 3.5×10^{-6}, respectively. The parameters λ in Eq. 6 is set to 0.01.

C. Ablation Studies

1) Quantitative analysis: To verify the effectiveness of each component and to determine their optimal combination, we conduct several ablation studies and the quantitative performance of each module is shown in TABLE I. It can be seen that compared with the baseline model, the performance of the model that incorporates the attention mechanism is improved in all cases. When spatial attention is employed alone (S), it achieves an increase of 1.0%/1.1%/1.4% in Rank-1 and 1.6%/2.2%/1.4% in mAP using the Market-1501, DukeMTMC-ReID and MSMT17 datasets, respectively. When channel attention is employed alone (C), the Rank1/mAP increases by 1.2%/1.7% using the Market-1501, 0.9% /2.4% using the DukeMTMC-ReID and 1.9%/1.8% using the MSMT17 compared with the Baseline. This demonstrates the importance of supervised attention for improving model performance.

Further, we place the two attention modules in a sequential (S-C, C-S) or parallel (S*C) manner. As can be seen from

TABLE I
ABLATION STUDY OF DIFFERENT ATTENTION FORMS. THE BEST PERFORMANCES ARE HIGHLIGHTED BY BOLD FONT

Method	Market-1501		DukeMTMC		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Baseline	93.2	82.6	88.1	75.3	76.1	51.2
S	94.2	84.2	89.2	77.5	77.5	52.6
C	94.4	84.3	89.0	77.7	78.0	53.0
S-C	94.0	84.5	89.0	77.1	78.3	52.7
C-S	93.9	84.7	88.8	77.4	78.5	52.7
S*C	**94.7**	**85.1**	**89.9**	**78.8**	**79.3**	**53.7**

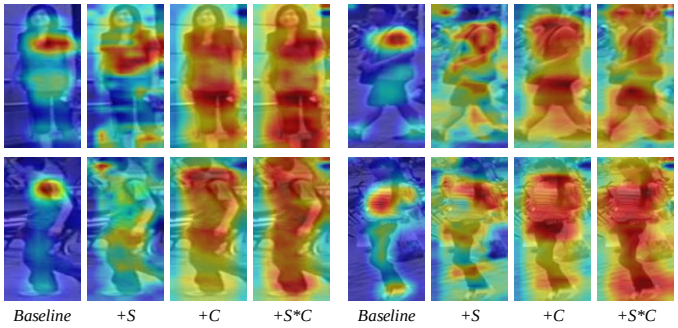

Fig. 2. The CAM produced by baseline, and the networks with spatial attention(S), channel attention(C), two attentions in parallel(S*C), respectively.

the TABLE I, generating an attention map sequentially does not bring significant performance, either in channel-first or spatial-first order. This may be due to the fact that the channel attention is obtained by the similarity between feature maps and spatial attention, and the computation of the two attentions in series is redundant. However, a parallel arrangement gives a better result than all the above methods. It exceeds the baseline by 1.5%/1.8%/3.2% in Rank-1 and 2.5%/3.5%/2.5% in mAP using the three datasets, respectively. This indicates that the two types of attention further enhance the ability of the model to extract discriminative features after multiplying them together.

2) Qualitative analysis: Fig. 2 shows the CAM output from the baseline and other attention networks. It reveals that the baseline focuses on the shoulders of the pedestrians, with slight attention to the lower extremities, ignoring the information of most regions of the pedestrian's body. The model with spatial attention only has paid more attention to the pedestrian region, but it's not a good output. The model can already focus on almost all body parts of the pedestrians after adding channel attention only. Furthermore, after using two kinds of attention simultaneously, many body parts change from yellow to red in CAM, and the model's ability to spotlight pedestrian body regions is further enhanced.

In addition, we intuitively reflect the retrieval performance of the model through the rank list of query. As illustrated in Fig. 3, the retrieval result of the baseline approach has eight error results. It is significantly reduced to three after adopting the supervised spatial attention, and further reduced to one after employing the supervised channel attention which in the last position of the retrieval results. The full-right retrieval results were achieved by applying a combination

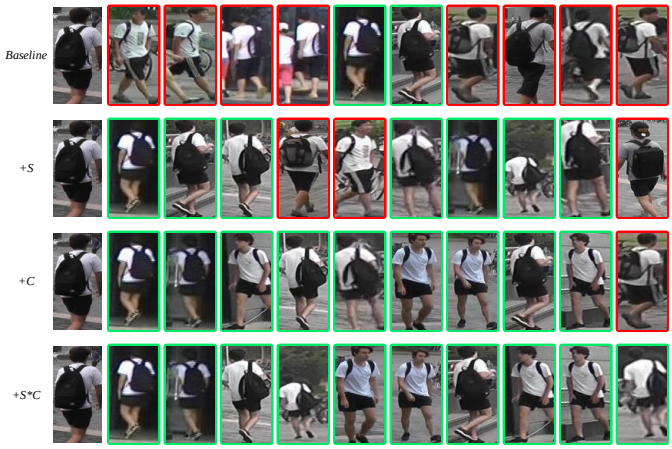

Fig. 3. The Rank-list produced by baseline, and the networks with spatial attention(S), channel attention(C), two attentions in parallel(S*C), respectively.

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON MARKET-1501, DUKEMTMC-REID AND MSMT17. (BOLDFACE DENOTES THE BEST RESULT, -: NOT AVAILABLE)

Method	Market-1501		DukeMTMC		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
HA-CNN [1]	91.2	75.7	80.5	63.8	-	-
CASN [2]	94.4	82.8	87.7	73.7	-	-
AANet [3]	93.9	83.4	87.7	74.3	-	-
IANet [23]	94.4	83.1	87.1	73.4	75.5	46.8
PCB+RPP [24]	93.8	81.6	83.3	69.2	68.2	40.4
VPM [25]	93.0	80.8	83.6	72.6	-	-
Li *et al.* [26]	90.2	82.7	81.0	78.0	-	-
ResNet50+CE-SAN [8]	94.1	84.1	84.8	74.2	77.3	**55.0**
PFFN [27]	**95.1**	84.6	88.9	78.7	72.9	48.2
Ours	94.7	**85.1**	**89.9**	**78.8**	**79.3**	53.7
Wang *et al.* [12]	94.7	91.7	89.0	85.9	-	-
Ours+re-rank	**95.1**	**92.7**	**90.1**	**86.4**	-	-

of the two attention approaches, illustrating the performance improvement of our proposed attention network for pedestrian re-identification.

D. Comparison With the State-of-the-Art Methods

The performance of the proposed approach is compared with state-of-the-art methods using three datasets in TABLE II. Compared with existing attention-based and part-based methods, our propsed method exceeds the best state-of-the-art method when using the DukeMTMC-reID, and achieves competitive results when using Market-1501 and MSMT17. Comparing with the current parameter-free attention method [12], we use re-rank [28] to test using the Market1501 and DukeMTMC-reID datasets, as this strategy is used in the literature. The Rank1/mAP achieves an increase of 0.3%/1.0% and 1.1%/0.5% using the Market1501 and DukeMTMC-ReID datasets respectively. This demonstrates that our proposed method still achieves good performance without training with additional parameters.

IV. CONCLUSION

We propose a CAM-guided Parameter-free Attention Network for person Re-ID. In contrast to the existing supervised

spatial attention, we use CAM as the supervised signal to directly guide the generation of attention in space. Different from the current unsupervised channel attention approaches, we design a non-linear approach to establish the association between the pedestrian identity and the channel. In addition, our attention models contain no training parameters and are only inserted once in the network, reducing the computational effort and the risk of model over-fitting. Experiments on Market1501, DukeMTMC-reID and MSMT-17 show that the proposed method is effective and achieves a competitive performance.

REFERENCES

[1] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogniti.*, 2018, pp. 2285–2294.

[2] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogniti.*, 2019, pp. 5735–5744.

[3] C.-P. Tay, S. Roy, and K.-H. Yap, "Aanet: Attribute attention network for person re-identifications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogniti.*, 2019, pp. 7134–7143.

[4] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "Abd-net: Attentive but diverse person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8351–8361.

[5] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 371–381.

[6] B. N. Xia, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3760–3769.

[7] G. Chen, C. Lin, L. Ren, J. Lu, and J. Zhou, "Self-critical attention learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9637–9646.

[8] X. Zhu, J. Qian, H. Wang, and P. Liu, "Curriculum enhanced supervised attention network for person re-identification," *IEEE Signal Processing Letters*, vol. 27, pp. 1665–1669, 2020.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogniti.*, 2016, pp. 2921–2929.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogniti.*, 2018, pp. 7132–7141.

[11] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," *ArXiv:1910.03151*, 2019.

[12] H. Wang, Y. Fan, Z. Wang, L. Jiao, and B. Schiele, "Parameter-free spatial attention network for person re-identification," *ArXiv:1811.12150*, 2018.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogniti.*, 2016, pp. 770–778.

[14] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogniti. Workshops*, 2019.

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 448–456.

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. on Comput. Vis.*, September 2018, pp. 3–19.

[17] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogniti.*, 2019, pp. 1389–1398.

[18] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[19] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *ArXiv:1703.07737*, 2017.

[20] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, p. 1116–1124.

[21] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. on Comput. Vis.*, 2016, pp. 17–35.

[22] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogniti.*, 2018, p. 79–88.

[23] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogniti.*, 2019, pp. 9317–9326.

[24] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. on Comput. Vis.*, 2018, pp. 480–496.

[25] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogniti.*, 2019, pp. 393–402.

[26] Y. Li, X. Jiang, and J.-N. Hwang, "Effective person re-identification by self-attention model guided feature learning," *Knowledge-Based Systems*, vol. 187, p. 104832, 2020.

[27] Y. Hou, S. Lian, H. Hu, and D. Chen, "Part-relation-aware feature fusion network for person re-identification," *IEEE Signal Processing Letters*, vol. 28, pp. 743–747, 2021.

[28] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogniti.*, 2017, pp. 1318–1327.