Self Attention based multi branch Network for Person Re-Identification

Self Attention based multi branch Network for Person Re-Identification

Asad Munir
Dept. of Computer Science
University of Udine
Udine, Italy
asad.munir@uniud.it

Christian Micheloni
Dept. of Computer Science
University of Udine
Udine, Italy
christian.micheloni@uniud.it

***Abstract*—Recent progress in the field of person re-identification have shown promising improvement by designing neural networks to learn most discriminative features representations. Some efforts utilize similar parts from different locations to learn better representation with the help of soft attention, while others search for part based learning methods to enhance consecutive regions relationships in the learned features. However, only few attempts have been made to learn non-local similar parts directly for the person re-identification problem. In this paper, we propose a novel self attention based multi branch(classifier) network to directly model long-range dependencies in the learned features. Multi classifiers assist the model to learn discriminative features while self attention module encourages the learning to be independent of the feature map locations. Spectral normalization is applied in the whole network to improve the training dynamics and for the better convergence of the model. Experimental results on two benchmark datasets have shown the robustness of the proposed work.**

***Index Terms*—person re-identification, self attention, multi branch network**

I. Introduction

The task of person re-identification (re-id) requires to retrieve a person's images from a gallery set gathered from non-overlapping cameras, given a query image. Person re-identification has gained progressive importance in the field of computer vision due to its applications in intelligent surveillance systems. Variations with respect to illumination, occlusion, resolution, viewpoints, pose, clothing and background in the images makes re-id a challenging task. With the rapid advancement of deep learning, ConvNets [1]–[3] well designed for image classification tasks have also provided a strong representatios of person's features for person re-id. Features from these networks outperform the traditional handcrafted low level features [4]–[7] by a large margin. However, there is a difference in image classification and person re-id tasks. This is that in re-id training and testing classes (i.e person identities) are not same. So person re-id requires more discriminative feature representations to overcome the problem of unseen classes (identities) at testing time.

Along with learning discriminative descriptor for person re-id, many existing studies [8]–[10] are also focusing on designing a better metric learning loss functions, including triplet loss, triplet hard loss, quadruplet loss etc. The purpose of these loss functions is to increase the generalization capability by reducing intra-class and enlarging inter-class variations. The issue with metric learning losses is that they are highly dependent on sample mining techniques and aren't always robust when there are outliers in the training set. On the contrast to metric learning methods, there are techniques which address the person re-id problem as classification task and compute cross entropy softmax classification loss for person identities. In testing stage, classification based approaches need to calculate distance matrix of features to distinguish different persons images. In many approaches, metric learning loss performs better than classification loss because of the mismatch between training and testing identities in person re-id task. To solve this issue, we propose a multi branch (classifier) network which allows the network to learn the most discriminative and robust features for each identity. The training with multiple classifiers enhances the re-id performance.

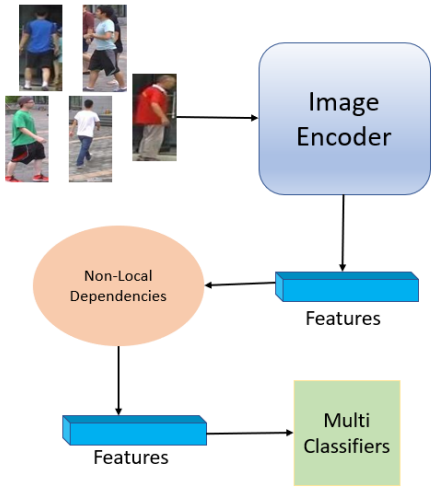

Fig. 1. Overall framework of the proposed method. ResNet50 is used as image encoder.

Recently, part based methods [11]–[13] have contributed towards learning part-informed representations of persons and

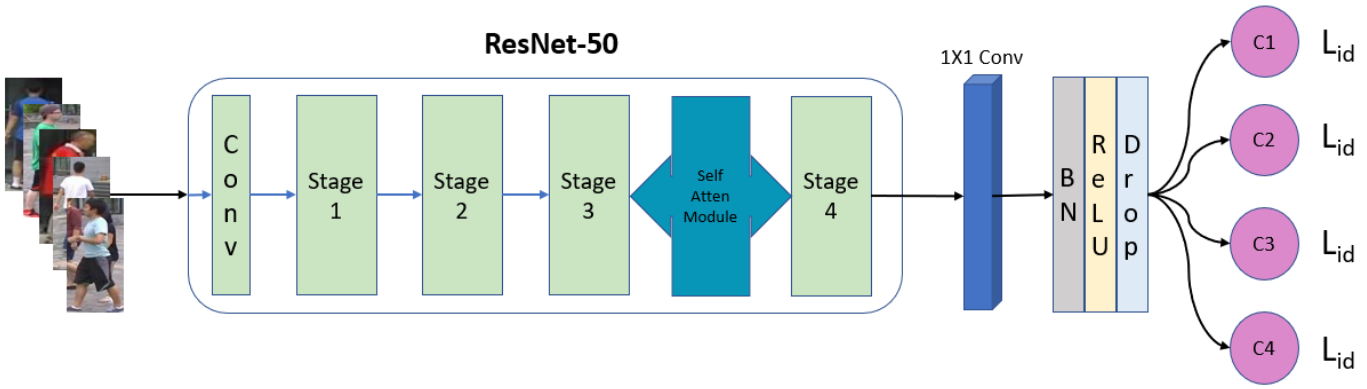

Fig. 2. Overview of the proposed framework. $C1$, $C2$, $C3$ and $C4$ are four fully connected layers for the predictions of person identity and their output losses are added to obtain the final loss. BN, Drop represent Batch-Normalization and Dropout Layers respectively.

achieved very promising performance for person re-id. These methods learn fine-grained features for each local part by splitting the backbone network's feature maps into horizontal local parts. The drawback of these models is that they need well aligned body parts of the same person to learn part based features and heavily rely on convolution to obtain the dependencies across different image regions. To process long range dependencies, several convolutional layers are required because convolution operator has a local receptive field. Due to which small models are unable to get these long range dependencies while large models have high computational costs. On the other hand, self attention [14] has better trade off between capturing long range dependencies and having reasonable computational ability. The self attention calculates the weighted sum of all features at all positions to build the response at current position and prevents high computational cost.

In this paper, we introduce self attention module to model long range dependencies which helps to emphasize similarities at different positions in the backbone network for person re-id. Along with multi classifier training, the addition of self attention module encourages the network to capture discriminative and robust person feature representations. The proposed framework significantly enhance the person re-id performance and is shown in Fig 1.

The main contributions in this work are given as follows:

- A multi branch (classifier) network to learn most discriminative features representations for person images to overcome the issue of mismatching identities in training and testing stage.
- The addition of a self attention module in the backbone network to model long range dependencies for finding similarities between different locations in the learned features.

With the addition of above mentioned contributions, We perform experiments on two benchmarks of person re-id. Results on these datasets show the performance and robustness of the proposed technique.

II. Self Attention Based Multi Branch Network

A. Problem definition and Notations

Let a set of n training images $\{I_i\}_{i=1}^{n}$ with corresponding identities labels $\{y_i\}_{i=1}^{n}$ be acquired by a camera network. Person re-identification requires to retrieve images of the same identity from a gallery set of different cameras by giving a probe image. The task of person re-id is very similar to image classification task when using cross entropy classification loss having different person identities in training and testing stage. With the help of a classifier, most disciminative features are learned to distinguish between two identities and these features are used to compute the distance matrix between the identities to perform person re-identification.

B. Proposed Architecture

Recent works have shown that Convolutional Neural Networks (CNNs) are efficient for learning deeper and robust feature representations from images and are accurate to train if they have shorter connection between layers. Relying on such outcomes, we define ResNet-50 [15] as our backbone network with several adjustments. We modify the stride (stride=1) of last downsampling block to make the spatial size of convolutional features larger before global average pooling by following the work of R-FCN [16]. We apply global max pooling instead of global average pooling on the features from the last downsampling block. A 1×1 convolution layer is added after the pooling to reduce the size of features channels from 2048 to 1024. This added convolution layer learns the most discriminative features from a person image. These features are then sent through batch normalization, Rectified Linear Unit (ReLU) and dropout layers and finally passed to multiple fully connected layers (classifiers) to predict the identity of the person in the input image. Since the gradients from all classifiers are gathered at previous convolution layer, thus they force that layer to learn the most discriminative global features

for computing distance matrix in testing stage. The learned global features are depending on local neighbourhood as the convolution layers have local receptive field. To learn the long-range dependencies, we add a self attention module in the backbone network to model these long-range dependencies and to capture the similar parts at different regions in the image. We append self attention block at the end of stage 3 as ResNet-50 consists of four stages after the first convolution block. The details of the self attention block is discussed in the next section. The proposed network is shown in Fig 2 and it is trained by using cross entropy loss which is given as:

$$L_{id} = -\sum_{c=1}^{C} \log(p(c))q(c) \tag{1}$$

where $p(c)$ is the output probability of the input belonging to class c. C, $q(c)$ are the total number of classes (person identities) in the dataset and ground truth distribution respectively. In the testing stage, the features from the last added 1×1 convolution layer are used to compute distance matrix.

C. Self-Attention module

Most of the person re-id models are built using convolutional layers. Due to the fact that the convolution processes the information in the local neighbourhood, convolutional layers are computationally unable to grasp long-range decencies in images. In the proposed method, we adapt a non local model [17] to introduce self-attention in a convoultional framework for the association of widely separated spatial regions.
From the previous hidden layers, the image features $x \in R^{C \times N}$ $(N = W \times H)$ are first modified into two feature spaces f, g such that $f(x) = W_f x$ and $g(x) = W_g x$ to compute the attention. $s_{ij} = f(x_i)^T g(x_j)$ and attention map $\beta_{j,i}$ is calculated as:

$$\beta_{j,i} = \frac{exp(s_{ij})}{\sum_{i=1}^{N} exp(s_{ij})} \tag{2}$$

where $\beta_{j,i}$ represents to which extant the model takes part in i^{th} location when synthesizing j^{th} region. Here, C, N are number of channels and number of feature locations of previous layer's features. The outputs of attention layer is $a = (a_1, a_2, ..., a_j, ..., a_N) \in R^{C \times N}$, and,

$$a_j = v\left(\sum_{i=1}^{N} \beta_{j,i} h(x_i)\right) \tag{3}$$

where $h(x_i) = W_h x_i$ and $v(x_i) = W_v x_i$. The formulation in eq 3 has learned weight matrices $W_f \in R^{\bar{C} \times C}$, $W_g \in R^{\bar{C} \times C}$, $W_h \in R^{\bar{C} \times C}$, and $W_v \in R^{\bar{C} \times C}$ which are implemented using 1×1 convolutions. $\bar{C}$ are the number of channels after reduction C/k, where $k = 1, 2, 4, 8$ and we are using $k = 8(i.e., \bar{C} = C/8)$ in our experiments for memory efficiency. For scaling, we multiply the output of the attention layer by a scale parameter and add back to the input feature map. The final output is given by

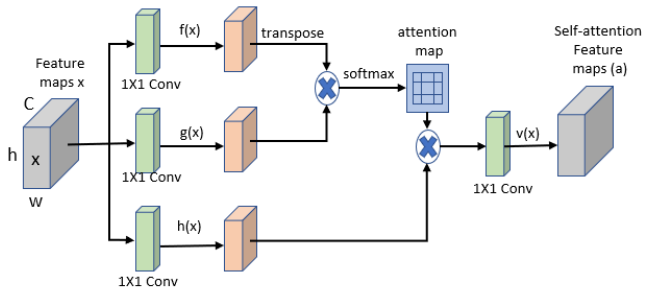

Fig. 3. Self attention module which is added after stage 3 in ResNet-50. The dimension of the output (self-attention) features is same as input because they are the input to stage 4 of the ResNet-50 Network.

$$y_i = \gamma a_i + x_i \tag{4}$$

where γ is learnable scalar parameter and is initialized as 0. γ is encouraging the network to rely first on the cues in the local neighbourhood and then progressively assign more weight to the non-local evidence. We apply self attention module after the second last stage of the backbone network as shown in Fig 3.

D. Spectral Normalization

Miyato et al. [18] originally proposed spectral normalization in the discriminator of Generative Adversarial Networks (GANs) [19] to stabilize the training which bounds the Lipschitz constant of the network by restricting the spectral norm of each layer. Spectral normalization does not need extra hyper parameter tuning like the other normalization techniques and also have relatively low computational cost. We also apply spectral normalization at every convolutional layer in the network to stabilize the training process without which the network starts diverging after few epochs.

III. Experiments

A. Datasets

We perform our experiments on two person re-id benchmark datasets the market-1501 [20] and the Duke-MTMC [21].
Market-1501 consists of 1501 identities automatically detected from six cameras. The query set has 3368 while gallery set has 19732 images with 750 identities. The training set has 12936 images of 751 identities.
Duke-MTMC dataset is composed of 36411 image of 1404 identities shot by eight cameras. There are 16522 images of 702 identities in training set. Query and testing set contains 2228 and 17661 images of 702 identities respectively.

B. Implementation Detail

We implemented the proposed network using Pytorch. The backbone network contains ResNet-50 network pretrained on ImageNet dataset with the modifications mentioned in section $II - B$. The network is optimized by using stochastic gradient Descent (SGD) with momentum 0.9 and the batch size is set

to 64. The initial learning rates for backbone network and layers added at the end are 0.001 and 0.01 respectively. We trained our model for 260 epochs in total with learning rate divided by 10 after 160 epochs. All the images are resized to 256×128 with random horizontal flipping and random erasing data augmentations. The dropout probability is 0.5 and weight decay is set to $1e-5$.

TABLE I
COMPARISONS TO THE STATE-OF-THE-ART RE-ID METHODS ON MARKET-1501. THE TOP 1 AND 2 RESULTS ARE IN RED AND BLUE.

Methods	Reference	Market-1501		
		Rank-1(%)	*Rank-5(%)*	*mAP(%)*
SpindleNet [22]	CVPR17	76.9	91.5	-
Part-Aligned [23]	ICCV17	81.0	92.0	63.4
HydraPlus-Net [24]	ICCV17	76.9	91.3	-
LSRO [25]	ICCV17	84.0	-	66.1
SVDNet [26]	ICCV17	82.3	92.3	62.1
DPFL [27]	ICCV17	88.9	92.3	73.1
PSE [28]	CVPR18	87.7	94.5	69.0
HA-CNN [29]	CVPR18	91.2	-	75.5
MLFN [30]	CVPR18	90.0	-	74.3
DuATM [31]	CVPR18	91.4	97.1	76.6
DKP [32]	CVPR18	90.1	96.7	75.3
GCSL [33]	CVPR18	93.5	-	81.6
PCB [13]	ECCV18	92.3	97.2	77.4
IDCL [34]	CVPRW19	93.1	-	78.9
CASN(IDE) [35]	CVPR19	92.0	-	78.0
SFT [36]	ICCV19	93.4	97.4	82.7
Proposed	-	93.8	97.4	80.8

C. Comparison with state-of-art methods

Table I shows the result of proposed method and its comparison with other state of art methods on market-1501 dataset. Highest results are shown in red while the second highest in blue. The proposed method have highest rank-1 accuracy while in terms of mean average precision (mAP) our methods produce comparable results.

In table II, results on Duke-MTMC dataset are presented with its comparison with other state of art methods. The Rank-1 accuracy and mean average precision of the proposed method is second highest with a very small difference and we obtain highest results in case of rank-5 accuracy. In all experiments, we reported our results while those of other methods are taken directly from the papers.

The results of the proposed method are considered without any re-ranking.

D. Ablation Study

In ablation study, we perform components analysis on the proposed network using market-1501. Firstly, we have shown the effect of multi classifiers training as compared to the single classifier. Only one loss is calculated when using single classifier while multiple losses are added to get the final loss in case of multi classifiers. We use 4 classifiers at the end of the proposed network. Multi classifiers outperform the single classifier marginally as shown in table III (first and second row). In the second phase, we place our self attention block at several positions in the backbone network and record its performance in table III. Backbone (ResNet-50) network is

TABLE II
COMPARISONS TO THE STATE-OF-THE-ART RE-ID METHODS ON DUKE-MTMC DATASET. THE TOP 1 AND 2 RESULTS ARE SHOWN IN RED AND BLUE.

Methods	Reference	Duke-MTMC		
		Rank-1(%)	*Rank-5(%)*	*mAP(%)*
Verif-Identif [37]	TOMM18	68.9		49.3
LSRO [25]	ICCV17	67.7	-	47.1
SVDNet [26]	ICCV17	76.7	86.4	56.8
DPFL [27]	ICCV17	73.2	-	60.6
PSE [28]	CVPR18	79.8	89.7	62.0
HA-CNN [29]	CVPR18	80.5	-	63.8
AACN [38]	CVPR18	76.8	-	59.2
MLFN [30]	CVPR18	81.0	-	62.8
DuATM [31]	CVPR18	81.8	90.2	68.6
DKP [32]	CVPR18	80.3	89.5	63.2
GCSL [33]	CVPR18	84.9	-	69.5
PCB [13]	ECCV18	81.8	-	66.1
IDCL [34]	CVPRW19	83.9	-	68.2
CASN(IDE) [35]	CVPR19	84.5	-	67.0
Proposed	-	84.6	91.6	68.6

composed of 4 stages and initial convolution block. All the four stages have 3, 4, 6 and 3 resisual blocks and we put the self attention module at the output of every stage. Among all other places, self attention block performance is higher when we place it at the end of stage 3 of ResNet-50.

TABLE III
ABLATION STUDY ON THE PROPOSED NETWORK.

Network Components	Market-1501	
	Rank-1(%)	*mAP(%)*
backbone (single classifier)	89.0	70.2
backbone (multiple classifiers)	92.4	78.6
backbone (attention after stage-1)	92.8	79.6
backbone (attention after stage-2)	93.5	80.4
backbone (attention after stage-4)	92.2	79.0
Proposed (attention after stage-3)	93.8	80.8

IV. CONCLUSION

We propose a novel self attention based multi branch network for person re-identification. Multi classifier training learns most discriminative features of the given person image to overcome the identities mismatching in training and testing stage. We add self attention module at the end of stage three in backbone (ResNet-50) network which models long-range dependencies to capture similarities at different feature locations instead of local neighbourhoods. Spectral normalization is applied to stabilize training and to avoid divergence of the model. The proposed network has the capability to learn robust feature representations and performs better than other state of art methods on two person re-id benchmarks.

ACKNOWLEDGMENT

This work was supported by EU H2020 MSCA through Project ACHIEVE-ITN (Grant No 765866).

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[4] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by hpe signature," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 1413–1416.

[5] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," in *European conference on computer vision*, 2014, pp. 330–345.

[6] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3610–3617.

[7] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3318–3325.

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

[9] N. Martinel, G. Luca Foresti, and C. Micheloni, "Aggregating deep pyramidal representations for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[10] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 403–412.

[11] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[12] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8514–8522.

[13] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 480–496.

[14] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[17] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[18] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[20] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 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 *European Conference on Computer Vision*, 2016, pp. 17–35.

[22] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1077–1085.

[23] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3219–3228.

[24] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 350–359.

[25] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3754–3762.

[26] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3800–3808.

[27] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2590–2600.

[28] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. 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*, 2018, pp. 420–429.

[29] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2285–2294.

[30] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2109–2118.

[31] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. 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*, 2018, pp. 5363–5372.

[32] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep kronecker-product matching for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6886–6895.

[33] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep crf for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8649–8658.

[34] Y. Zhai, X. Guo, Y. Lu, and H. Li, "In defense of the classification loss for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[35] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 5735–5744.

[36] C. Luo, Y. Chen, N. Wang, and Z. Zhang, "Spectral feature transformation for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4976–4985.

[37] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1, pp. 1–20, 2017.

[38] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2119–2128.