Further Non-local and Channel Attention Networks for Vehicle Re-identification

Kai Liu, Zheng Xu, Zhaohui Hou, Zhicheng Zhao, Fei Su
Beijing University of Posts and Telecommunications, Beijing, China
Beijing Key Laboratory of Network System and Network Culture, Beijing, China
{lk_dyx, xuzheng, doffe, zhaozc, sufei}@bupt.edu.cn

Abstract

Vehicle re-identification remains challenging due to large intra-class difference and small inter-class variance. To address this problem, in AICity Vehicle Re-ID task 2020, we propose a two-branch adaptive attention network—Further Non-local and Channel attention (FNC) to improve feature representation and discrimination. Specifically, inspired by two-stream theory of visual cortex, based on Non-local and channel relation, a two-branch FNC network is constructed to capture multiple useful information. Second, an effective attention fusion method is proposed to sufficiently model the effects from spatial and channel attention. The experimental results show that our algorithm achieves 66.25%/Rank-1 and 53.54%/mAP in 2020 AICity Challenge Vehicle Re-ID task without using extra data, annotation and other auxiliary information, which demonstrate the effectiveness of the proposed FNC network.

1. Introduction

Vehicle re-identification (Re-ID) refers to the recognition of the car of interest in different cameras. This process possesses diverse real-world applications and plays an important role in AI cities, which are now attracting increasing attention. However, conducting vehicle Re-ID is challenging when large intraclass variants (e.g., viewpoints, illuminations, and occlusions) are present.

Given that person and vehicle Re-ID conceptually belong under image retrieval problems, some commonly used strategies for the former are also useful for the latter. Most of state-of-the-art CNN-based person Re-ID methods adopt pretrained CNN models (e.g., ResNet [5]) on ImageNet and fine-tune them on the Re-ID datasets under the supervision of different losses (e.g., softmax and triplet losses [6]). In person Re-ID, the human body is vertically symmetrical and can be partitioned into head, torso, legs, and feet along the height dimension, rendering height-wise partition practical[12][16][21]. However, this type of partition is not sufficient in vehicle Re-ID because predicting the direction of the vehicle is difficult. Therefore, this study focuses on enhancing the representation of the global features instead of performing a height-wise partition.

Attention mechanism plays an important role in the human visual perception system. This system is widely used in Re-ID because of its capability to let the model focus on the subject of the target rather than the background. From a dimensional perspective, attention mechanism can be divided into two main categories. The first category is spatial attention, which focuses on “where” the informative parts of a given image are (e.g., Non-local block [17]). The second one is channel attention, which determines “what” the most meaningful parts of an image are (e.g., SENet [7]). Several recent works have combined the two categories into one attention model (e.g., CBAM [19]).

However, modeling visual attention is still insufficient in most existing works. The attention network like Non-local block may be further explored to enhance its suitability for the vehicle Re-ID. In this paper, we propose a simple yet effective approach, Further Non-local and Channel attention (FNC), to effectively learn the discriminative features and eliminate the negative impacts caused by the background of the vehicle image. The proposed method aims to simultaneously utilize the global and channel information of a vehicle image for the Re-ID task in a highly robust and efficient manner. The contributions of this paper are summarized as follows.

1) A two-branch adaptive attention network, i.e., Further Non-local and Channel attention (FNC) is constructed to simulate two-stream theory of visual cortex, and additionally, empirical network architecture and training strategy are explored and compared.

2) Based on Non-local and channel relation, two blocks, namely, spatial attention block (SAB) and channel attention block (CAB) are built and fused by a sigmoid function to emphasize spatial and channel attention, which enhances feature discrimination.

3) The proposed algorithm achieves 66.25%/Rank-1 and 53.54%/mAP in the AI City 2020 Challenge vehicle

Re-ID task without using any additional data, annotation and other auxiliary information.

2. Related works

The main purpose of Re-ID is to facilitate the network in extracting a discriminative feature that can accurately represent the image. Related studies can be roughly summarized into three aspects.

Receptive field optimization: Most of the feature extraction networks are implemented on the basis of CNN; thus, the features are mainly extracted from convolution operations. The filter can only process one local neighborhood at a time due to the limitation of the kernel size. If the filter can obtain the global receptive field, then the network can achieve an enhanced performance. Wang et al. [17] proposed a Non-local structure that compromised some ideas of spatial attention. The Non-local block generates a weighted mask relies on the similarity among mapped pixels, and it then computes the response at a position as a weighted sum of the features at all positions. This process allows the distant pixels to contribute to the filtered response at a location based on the patch appearance similarity, which yields a global receptive field. The SENet proposed in [7] is a channel-wise attention mechanism. Two FC layers are connected after the global average pooling of each channel. The structure then learns how to generate a weighted mask that can place large attention on the right channel of the feature map for each identity and establishes the connection for each channel. These receptive field optimizations promote several improvements in recent tasks and competitions.

Local feature extraction: Given that the scale of Re-ID datasets is not always extremely large, training using the entire image may cause the network to acquire the background information of the dataset, which always leads to overfitting. To address this problem and improve the generalization capability of the Re-ID models, Wang et al. [16] designed a multiple granularity network (MGN), which consists of one branch for global feature representations and two branches for local ones. The images are divided into several stripes, and these branches can obtain the local feature representations with multiple granularities. This structure forces each branch to learn from their local parts. Similar ideas have also been presented in other studies. The batch feature erasing (BFE) approach [1] introduces a "feature dropping branch," in which some parts of the feature map will be cropped. In [24], the cropped operation is implemented on the image, and some image parts are randomly deleted. To some extent, these methods force the network to learn further about the identity than the background, thereby preventing overfitting.

Loss optimization: Most works have combined ID and triplet loss to constrain the same feature. ID loss aims to constructs several hyperplanes to separate the embedding space into different subspaces, in which the cosine distance is more suitable than the Euclidean one. Conversely, triplet loss aims to enhance the intraclass compactness and interclass separability in the Euclidean space and then selects the Euclidean distance. If both losses will be used to simultaneously optimize a feature vector, then the goals may be inconsistent, and the network might hardly converge. To overcome this problem, Luo et al. [10] designed a structure called BNNeck, which adds a batch normalization (BN) layer after the origin feature. In this case, triplet loss will optimize the feature generated before the BN layer, and the ID loss will be computed by the feature made by the BN layer. This structure eases the conflict of the two losses, as well as the separation of the embedding space into different subspaces.

Although above mentioned works use attention mechanism to learn weight maps, not all weight maps play a positive role on feature representation. Moreover,[2] propose to partition the image into height-channel and width-channel to obtain local feature. This scheme is effective, however, it significantly increase computational cost. [8] use key-point model to infer vehicle orientation and use vehicles semantic attributes to help improve performance. Attributes are highly useful for Re-ID , yet they require additional annotation and other pretrained model. Therefore, we propose a novel two-branch Re-ID framework, in which visual attention is explored sufficiently, and comparable results are achieved without using any additional data, annotation and other auxiliary information.

3. Proposed method

This section describes the proposed network. FNC consists of a backbone architecture similar to the one used in [10], a proposed spatial attention module based on the Non-local block and a modified channel attention module.

3.1. Network architecture

Figure 1 shows the overall network architecture, which includes a backbone network, a spatial attention branch, and a channel attention branch. We use SE-ResNeXt-101 as the building foundation to enhance the feature extraction capability of the backbone network. Then, we change the last spatial down-sampling operation stride from 2 to 1 to provide a large spatial view for the spatial attention module, thereby capturing highly detailed spatial correlations. This manipulation causes a minimal increase in the computation cost and does not involve additional training parameters. We follow the modification strategies in state-of-the-art person Re-ID models [12] [16]. Accordingly, we duplicate the convolutional layers after the conv4_1 layer to split the SE-ResNeXt-101 into two branches. Finally, we adopt the BNNeck [10] to separate the optimized ID and triplet losses.

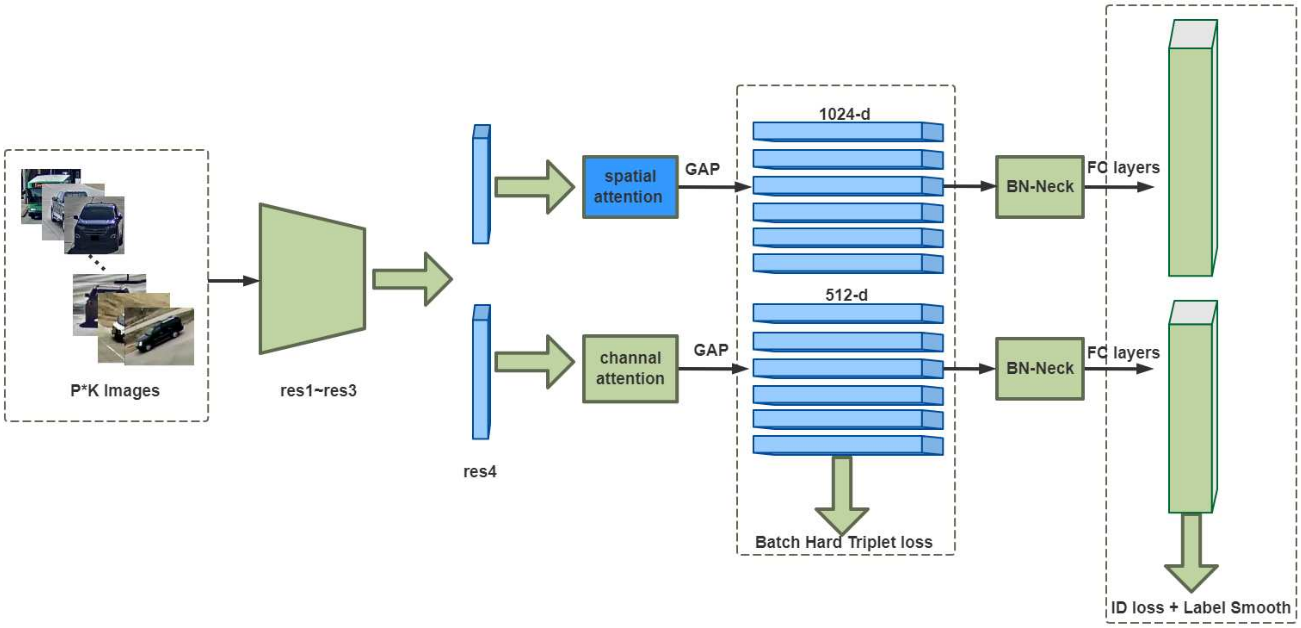

Figure 1. Overall network architecture, se-resnext-101 is used as the backbone network. Layers after res3 are duplicated to split our network into 2 independent branches. Spatial attention and channel attention are the corresponding SAB and CAB, and will be described in section 3.2 and 3.3.GAP refers to Global Average Pooling.BN-Neck is the BNN layers. In the inference stage, two branch feature vectors are concatenated as an appearance signature($dim = 1536$).

The spatial and channel attention branches share a similar structure; both branches consist of a global average pooling layer, a feature reduction module, and a BNNeck module. The global average pooling layer produces a 2048-dimensional vector from the spatial or channel attention module. The feature reduction module contains a 1*1 convolution, a BN layer, and a rectified linear unit (ReLU) layer to reduce the dimension to 1024 or 512, thereby providing a compact feature representation. We then use the BNNeck to separately normalize the feature for the triplet and ID losses.

3.2. The spatial attention block (SAB)

Figure2 shows the overview of the SAB module. Let $x \in R^{B*C*H*W}$ be the input to the SAB module, where B is the batch number; H and W are the spatial height and width of the tensor, respectively; and C is the number of the channels.

We use a 1*1 convolution in forming function g to reduce the number of channels C to C/r and reshape the tensor to $B * HW * C/r$. r refers to the reduction factor, which is set to 2 in our experiments, h and p play the same role as g. Therefore, we obtain two tensors with shape $HW * C/r$. Subsequently, we apply matrix multiplication to determine the Non-local relation and use softmax function to change the value to the probability of the region. After this operation, we can acquire an attention map x^p, which is then multiplied to x^r we get from function p to obtain the weighted feature map.

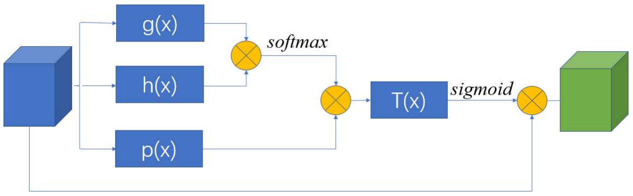

Figure 2. The architecture of the spatial attention block (SAB). $g(x)$ and $h(x)$ are the 1*1 convolution to reduce the dim from c to c/r. $p(x)$ plays the same role to adapt the attention weight. $T(x)$ is a simple learnable transformation.

$$x^p = softmax(g(x)h(x)) \tag{1}$$

The tensor shape is now $B * HW * C/r$. We use a simple learnable transformation T, which in this case is a 1*1 convolution, to restore the channel dimension of the attended tensor from C/r to C. We do not directly use element-wise addition to obtain the final feature map. Instead, we use a sigmoid function to activate the reshaped feature and apply element-wise multiplication to the origin feature map to obtain the final feature.

$$S_f = sigmoid(T(x^p * x^r)) * x^{raw} \tag{2}$$

By using the sigmoid function, the effect of the weight value on the feature map can be increased, and nonlinear factors can be introduced.

3.3. The channel attention block (CAB)

CAB aims to generate a channel mask to indicate the important channel responses, which serve as the supplement to the spatial branch. Figure 3 shows the CAB architecture, where SE-block is the squeeze and excitation module in [7]. The first step is to feed the raw feature map $x^f \in B * C * H * W$ into the two convolutional blocks.

$$C_f = sigmoid(SE(x^{raw})) * x^f \tag{3}$$

Each block consists of three consecutive operations: a convolutional layer, a BN layer, and a ReLU. The first convolutional block has 1024 filters and a kernel size of 1*1 to reduce the feature dimension. Similarly, the second convolutional block has 1024 filters, but the kernel size is set to 3*3. Moreover, the latter uses 32 groups to enhance the feature expression capability.

Next, the output of the two blocks are fed into another convolutional layer with a kernel size of 1*1 to generate the spatial map $x^{raw} \in B * C * H * W$. Then, we use the SE block to obtain the part channel attention. Finally, the channel attention map (i.e., SAB) is normalized into [0, 1] through the sigmoid function to multiply the origin feature map.

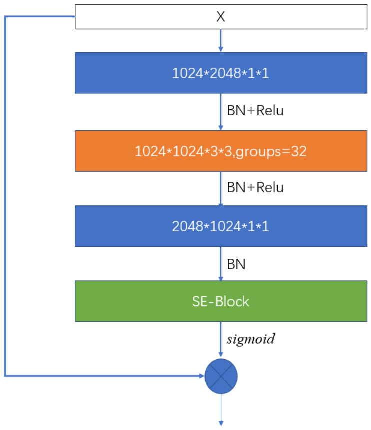

Figure 3. The architecture of the channel attention block (CAB). Four numbers refer to the output channel, the input channel and the shape of the kernel in sequence.

On the basis of previous works [1] [17], we directly added the attention map on the origin map to enhance the response of some regions. However, this step is cannot sufficiently reflect the weight function (the details for this claim will be discussed in Section 4.4). Therefore, we use the sigmoid function in the two branches to normalize the attention map to [0,1] and achieve an enhanced performance.

3.4. Loss functions

Cross entropy loss with smooth label. Cross-entropy loss is the most commonly used loss function in Re-ID tasks. To prevent the Re-ID model from overfitting the training IDs, we generate a soft label for each image [13] to facilitate the smooth training of the model.

The loss can be expressed as

$$L_{id} = \sum_{i=1}^{N_i} - q_i \, log\, (p_i) \tag{4}$$

where N_i denotes the number of images in the mini batch, p_i is the ID prediction logits of class i, and q_i can be defined as

$$q_i = \begin{cases} 1 - \frac{N-1}{N}\varepsilon, \; if \; y = 1 \\ \frac{\varepsilon}{N}, \; otherwise \end{cases} \tag{5}$$

where i is the index of the image; y is the identification of the image; N is the number of the samples in the dataset; and ε is a small constant, which is set to 0.1 in our implementation.

Triplet loss. The triplet loss with hard mining is introduced in [6] as an improved version of the original semi-hard triplet loss [11]. We randomly sample P identities and K instances for each mini batch to meet the requirement of the batch-hard triplet loss.

The triplet loss can be defined as

$$\begin{aligned} L_{triplet} = & \sum_{i=1}^{P}\sum_{a=1}^{K}[\alpha + \max_{p=1,...,K} ||a_i - p_i||_2 \\ & - \min_{n=1,...,K, j=1,...,P, j \neq i} ||a_i - n_j||_2]_+ \end{aligned} \tag{6}$$

Total loss. We utilize the batch-hard triplet and the softmax losses; each branch has one triplet loss and one softmax loss. Then, we determine the sum of the two branch losses.

$$L_{total} = \lambda \sum_{i=1}^{N} L_i^{id} + \sum_{j=1}^{N} L_j^{triplet} \tag{7}$$

where N is the branch number, and $\lambda = 2$.

4. Experiment details

	Training set	Validation set	Testing set
Identifies	233	100	333
Images	24627	12208	18290

Table 1. The information of the dataset.

4.1. Dataset and evaluation metrics

We conduct the experiments on the benchmark dataset [15] shown in Table 1, which contains 666 vehicles (56277 images in total) captured by 40 cameras in a real-world traffic environment. On average, each vehicle has 84.5 image signatures from 4.5 camera views. A total of 36935 images are included in the training set and 1052 images are used for the query in the testing set. We then divided the training set into a smaller training set and a validation set. The latter comprises 100 identities and 12208 images. After the appropriate hyperparameters are obtained, we train the network using the entire training set. The test set consists of 18,290 images belonging to the other 333 identities. The main evaluation metrics include the mAP and the Top-1/Top-5 accuracy of the CMC.

4.2. Implementation details

SE-ResNeXt101 [7] is chosen as the backbone to generate the feature. Given that a high spatial resolution can enrich the granularity of the feature, we remove the last spatial down-sampling operation in the backbone network. The RGB channels for each pixel are normalized, and the images are resized to 288*384 [14] before feeding to the network. Label smoothing and random erasing augmentation are also implemented to improve the generalization capability of the Re-ID model. The smoothing parameter ε and the probability of random erasing are set as 0.1 and 0.5, respectively.

The initial learning rates are set as $3.5*10^{-5}$. Ten epochs are used to linearly increase the learning rate to $3.5*10^{-4}$ as warm up [4]. Then, the learning rate is decayed to $3.5*10^{-5}$ and $3.5*10^{-6}$ at the 40th and 70th epochs, respectively. The batch size is set to 64 and the Ranger (i.e., combination of LookAhead [20] and RAdam [9]) is adopted as the optimizer. In the inference stage, we select the k-reciprocal re-ranking method [23] as the re-ranking algorithm. The hyperparameters are set as follows: $k1 = 75$, $k2 = 10$, and $\lambda = 0.1$.

4.3. Experimental results

We compare the proposed model with other state-of-the-arts Re-ID models on the validation dataset.

method	Top-1	Top-5	Top-10	mAP
Baseline(resnet-50)	78.0%	83.0%	87.0%	58.3%
Baseline(se-resnext-101)	81.0%	86.0%	88.0%	61.8%
MGN[16]	82.0%	87.0%	90.0%	65.4%
BFE[3]	84.0%	**91.0%**	**93.0%**	66.6%
FNC(ours)	**85.0%**	87.0%	91.0%	**70.8%**

Table 2. The table shows mAP and Top1/5/10 in our validation set. MGN and BFE both use se-resnext-101 as the backbone for fair comparison.

In the validation dataset, we select the strong baseline [10] due to its high performance on person Re-Id by only using the global features without any other additional information. This baseline is therefore suitable for person and vehicle Re-ID. In addition, different backbones may lead to different results [22]. As shown in Table 3, the SE-ResNeXt101 [7] achieves the most satisfactory performance in the validation set.

backbone	Top-1	mAP
Resnet-50	78.0%	58.3%
Resnet-101	79.0%	59.5%
Resnet-101-ibn-a	81.0%	61.3%
Se-Resnext-101	**81.0%**	**61.8%**

Table 3. The Top-1 and mAP accuracy with different backbones on the validation set.

Optimizing large neural networks and selecting them as the strong baseline's backbone network can be difficult. And in the experiments, the center loss [18] will degrade the performance, which is why this method is not used in the baseline. In comparison with the baseline (i.e., ResNet-50), the proposed model can achieve a 7%/12.5% advance on Top-1 and mAP. To compare MGN and BFE fairly, we use SE-ResNeXt-101 as their new backbones. Table 2 shows that BFE performs more satisfactorily in terms of the Top-5 and Top-10 than FNC. This result might be because the former can handle some extent occlusion better than FNC, and the latter is not sufficient when only random erasing is used. Therefore, we fuse the BFE result with our proposed model to improve the performance.

method	Top-1	mAP
DenseNet121+Xent+Htri[15]	51.7%	31.0%
ResNext101+Xent+Htri[15]	48.8%	32.0%
MobileNetV1+BH[15]	48.4%	32.0%
Baseline(se-resnext-101)[10]	60.74%	46.40%
BFE(se-resnext-101)[3]	58.75%	47.98%
FNC(separate)	62.93%	50.92%
FNC+BFE	**66.25%**	**53.54%**

Table 4. Comparison of results on the testset of CityFlow-ReID[15] with 2 evaluation metrics: Top-1 and mAP. FNC+BFE refers to FNC feature and BFE feature.

We also compare the results of the FNC with the several baselines mentioned in the CityFlow dataset [15]. The results based on a private test set are reported in Table 4. FNC outperforms the baselines on the Top-1 accuracy and mAP by a large margin. In comparison with the other methods in the 2019 AI City Challenge, we did not use any additional data nor annotate the training data and only adopt two model ensembles to obtain comparable results.

4.4. Ablation experiments

Extensive experiments are conducted on the validation dataset to verify the effectiveness of the crucial components in FNC. We compare the performance of the different structures to identify the optimal architecture for the proposed model.

method	Top-1	mAP
Baseline(se-resnext-101)	81.0%	61.8%
Baseline + CAB	82.0%	66.0%
Baseline + SAB	83.0%	68.6%
FNC(not separate)	**84.0%**	68.5%
FNC(separate)	**84.0%**	**70.8%**

Table 5. SAB and CAB are the spatial attention block and channel attention block. FNC(not separate) refers to use same res4 in two branch. The results are on the validation set.

SAB and CAB: We separately test the effects of SAB and CAB on the model's advance (Table 5). The results show that using only CAB can garner 1%/4.2% Top-1 accuracy and mAP advance. When only SAB is used, the model yields 2%/6.6% Top-1 accuracy and mAP advance. However, when both branches are used to train the network together, only 1%/0.1% increment from the results of using SAB is obtained. It seems that the CAB branch is not necessary for the whole network. However, the experiments suggest that CAB brings advance indeed. The problem is that the same res4 is used when we train the entire network.But when we form the spatial and channel attention maps, they may produce totally different effect on the res4, even the oppositely. We then separated res4 from res3 in Figure 1. We obtain the final version of FNC and achieve 3%/9% Top-1 accuracy and mAP advance compared with the baseline. On the private test set, we reach a 5.51%/7.14% advance compared with the baseline, as well as a large margin of the official baseline in Table 4.

method	Top-1	mAP
Baseline(se-resnext-101)	81.0%	61.8%
FNC(separate + add)	**84.0%**	67.1%
FNC(separate+ multiply)	**84.0%**	**70.8%**

Table 6. 'add' refers to add residual x with the processed feature to learn residual in [1][17]. 'multiply' refers to this paper method by using sigmoid to activate weight value. The results are on the validation set.

Addition or multiplication: In common practice [1][17], the residual x is added to the processed feature x^p to learn the residual and reduce the learning difficulty. However, this process will weaken the attention weight, and the model cannot focus on the important spatial or channel region. Therefore, we use a simple function (i.e., sigmoid function) to normalize the weighted value to [0,1]. Then, we multiply this value with the origin map, which enlarges the weighted value effect compared when the value is added. Table 6 shows that multiplication can obtain a 3.7% mAP advance compared with addition.

5. Conclusion

In this paper, inspired by typical attention model, used in person Re-ID, we propose an adaptive attention network for vehicle Re-ID. By obtaining Non-local and channel relation, we use sigmoid function to enlarge the weighted value and get more sufficient spatial and channel attention map. In the inference stage, we concatenate two part feature to obtain better performance. In 2020 AICity challenge track2 vehicle Re-ID task, our algorithm achieves 66.25%/Rank-1 and 53.54%/mAP without using extra data, annotation and other auxiliary information, which demonstrate the effectiveness of the proposed FNC network.

6. Acknowledgement

This work is supported by Chinese National Natural Science Foundation (U1931202).

References

[1] B. Bryan, Y. Gong, Y. Zhang, and C. Poellabauer. Second-order non-local attention networks for person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3759–3768, 2019.

[2] Hao Chen, Benoit Lagadec, and Francois Bremond. Partition and reunion: A two-branch neural network for vehicle re-identification. In *Proc. CVPR Workshops*, pages 184–192, 2019.

[3] Zuozhuo Dai, Mingqiang Chen, Siyu Zhu, and Ping Tan. Batch feature erasing for person re-identification and beyond. *arXiv preprint arXiv:1811.07130*, 1(2):3, 2018.

[4] Xing Fan, Wei Jiang, Hao Luo, and Mengjuan Fei. Sphereid: Deep hypersphere manifold embedding for person re-identification. *Journal of Visual Communication and Image Representation*, 60:51–58, 2019.

[5] 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*, pages 770–778, 2016.

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

[7] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[8] Tsung-Wei Huang, Jiarui Cai, Hao Yang, Hung-Min Hsu, and Jenq-Neng Hwang. Multi-view vehicle re-identification using temporal attention model and metadata re-ranking. In *Proc. CVPR Workshops*, pages 434–442, 2019.

[9] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

[10] 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*, pages 0–0, 2019.

[11] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[12] 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.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[14] Xiao Tan, Zhigang Wang, Minyue Jiang, Xipeng Yang, Jian Wang, Yuan Gao, Xiangbo Su, Xiaoqing Ye, Yuchen Yuan, Dongliang He, et al. Multi-camera vehicle tracking and re-identification based on visual and spatial-temporal features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 275–284, 2019.

[15] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8797–8806, 2019.

[16] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018.

[17] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[18] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[19] 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.

[20] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, pages 9593–9604, 2019.

[21] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8506–8514, 2019.

[22] Zhedong Zheng, Tao Ruan, Yunchao Wei, and Yi Yang. Vehiclenet: Learning robust feature representation for vehicle re-identification. In *Proc. CVPR Workshops*, 2019.

[23] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3652–3661, 2017.

[24] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.