Robust Person Re-Identification through Contextual Mutual Boosting

Zhikang Wang\textsuperscript{1,2*}, Lihuo He\textsuperscript{1}, Xinbo Gao\textsuperscript{1} and Jane Shen\textsuperscript{2}

\textsuperscript{1}Xidian University
\textsuperscript{2}Pensees Singapore
{zkwang00, lihuo.he}@gmail.com, xbgao@mail.xidian.edu.cn, jane.shen@pensees.ai

Abstract

Person Re-Identification (Re-ID) has witnessed great advance, driven by the development of deep learning. However, modern person Re-ID is still challenged by background clutter, occlusion and large posture variation which are common in practice. Previous methods tackle these challenges by localizing pedestrians through external cues (e.g., pose estimation, human parsing) or attention mechanism, suffering from high computation cost and increased model complexity. In this paper, we propose the Contextual Mutual Boosting Network (CMBN). It localizes pedestrians and recalibrates features by effectively exploiting contextual information and statistical inference. Firstly, we construct two branches with a shared convolutional frontend to learn the foreground and background features respectively. By enabling interaction between these two branches, they boost the accuracy of the spatial localization mutually. Secondly, starting from a statistical perspective, we propose the Mask Generator that exploits the activation distribution of the transformation matrix for generating the static channel mask to the representations. The mask recalibrates the features to amplify the valuable characteristics and diminish the noise. Finally, we propose the Contextual-Detachment Strategy to optimize the two branches jointly and independently, which further enhances the localization precision. Experiments on the benchmarks demonstrate the superiority of the architecture compared the state-of-the-art.

1 Introduction

Person Re-Identification (Re-ID) aims at retrieving images of the same person as a given query image from a ‘gallery’ set of images. Numerous methods [Qian et al., 2018; Si et al., 2018; Sun et al., 2018; Su et al., 2017; Song et al., 2018] have been proposed to address the problem due to the significant impact of Re-ID in real-life application. However, existing methods typically suffer from performance degradation due to challenging factors like occlusion, background clutter, and posture changes. Therefore, there is consequently an inevitable need for training a robust model that can localize and recognize the pedestrians accurately.

Localizing pedestrians in an image is an effective way to alleviate negative effects from the above distracting factors. Motivated by the idea, numerous methods [Zheng et al., 2019; Song et al., 2018; Zhao et al., 2017; Kalayeh et al., 2018a; Qi et al., 2018; Liu et al., 2018] have been developed to generate discriminative and robust person representations. Among them, some methods [Liu et al., 2018; Cho and Yoon, 2016] perform pose estimation at first for inferring the location of the pedestrians indirectly. Some other methods [Kalayeh et al., 2018a; Song et al., 2018] adopt auxiliary algorithms to segment pedestrians from the images directly. Differently, the attention mechanism, which is capable of localizing the informative parts of an input image, is also adopted for coarsely localizing the pedestrians and filtering distracting factors [Li et al., 2018]. However, all these methods suffer from a common drawback, i.e., the high computation cost for both the training and inference phases. Besides, the performance of those methods that require external clues (e.g., pose, segmentation) relies heavily on the accuracy of those external clues, hurting the robustness of the overall model.

Recalibrating the features is also effective in improving the discrimination of the final representations for recognition. As each feature channel is considered as a specific feature detector [Zeiler and Fergus, 2014], recalibrating the attention weights on the channel dimension can dilute the noise and amplify the valuable characteristics. Quite a few methods [Li et al., 2018; Woo et al., 2018; Hu et al., 2018] focus on this line. However, they all follow the ‘squeeze and excitation’ mode [Hu et al., 2018], that first produces the channel descriptor through a squeeze operation and then produces a collection of channel modulation weights from the embeddings through an excitation operation. The limitation is that only stacking many such blocks can bring significant improvements to the discrimination, incurring high computation overhead.

In this paper, we propose the Contextual Mutual Boosting Network (CMBN) to generate more discriminative features for pedestrians by spatial localization and channel recalibra-
Figure 1: The overall architecture of our proposed Contextual Mutual Boosting Network (CMBN). There are four components, i.e., the shared convolutional frontend, the Backbone Branch, the Reversal Branch, and the Mask Generator Model. GRL and FC refer to the Gradient Reversal Layer and Fully-Connected Layer, respectively. During training, three ID losses will jointly optimize the whole network. In the inference phase, to be computationally efficient, the Reversal Branch and the Mask Generator will be discarded. The features multiplied by the generated mask form the final representation.

In summary, the contributions of this paper are as follow:

- We propose the Contextual Mutual Boosting Network (CMBN), which localizes pedestrians and recalibrates features by exploiting the contextual information and statistical inference, respectively.
- We propose a novel Reversal Branch, which aims at extracting the background information of images to boost the pedestrians’ spatial localization indirectly.
- We propose a novel Mask Generator to learn a universal and static channel mask that will recalibrate the features on the channel dimension to amplify the valuable characteristics and diminish the noise.
- We propose a Contextual-Detachment Strategy to optimize the two branches with contrary functions. It helps the Backbone Branch with better foreground sensitivity and more accurate spatial localization.

To demonstrate the effectiveness of our approach, we conduct experiments on three challenging Re-ID datasets. Our method achieves competitive performance under multiple evaluation metrics.

2 Related Work

Deep Learning Based Person Re-ID

The deep learning has witnessed great success in person Re-ID. Many works emerge after deep learning surpassing human-level performance in image classification tasks like ImageNet [Deng et al., 2009]. Meanwhile, many large scale person Re-ID datasets that enable training of deep networks become available, for instance Market1501 [Zheng et al., 2015], CUHK03 [Li et al., 2014] and DukeMTMC-ReID[Gou et al., 2017], which accelerate the development in this area. In [Zheng et al., 2016], an identification network is proposed for tackling person Re-ID. Afterwards, some domain-specific designs emerged, like PCB[Sun et al., 2018], MGN[Wang et al., 2018a].

Human Body Information

In this work, we attempt to generate representations robust to occlusion, posture variations, and background clutter. Similar approaches make use of body key points, parts region, human parsing or segmentation. In [Zhao et al., 2017], the authors apply a coarse-to-fine region feature extraction and merging process to generate robust and discriminative person representations. In [Miao et al., 2019], a similar divide-then-merge method are exploited. The authors use key points to generate attention and guide the network to focus on areas that can be seen so that occluded parts are not influencing the representation. In [Qi et al., 2018], the authors propose to use human body segmentation mask to guide the network to learn robust features. Moreover, a fine-grained human body segmentation is proposed in [Kalayeh et al., 2018b] to generate features for each semantic part.
Attentions

Those aforementioned methods utilize auxiliary modules with fixed parameters. Such configurations are sub-optimal because of mismatch between the domains of the main model and the auxiliary model. Hence, end-to-end-trainable modules are in more favor (than those require multiple training stages), such as attention modules. Here, we elaborate on two kinds of attention mechanisms: spatial attention and channel attention.

Spatial attention is useful for allocating available resources towards the informative parts of images. For example, [Wang et al., 2018b] introduced spatial self attention on the feature maps before the Global Average Pooling. [Xu et al., 2018] aggregates the pose estimation with the spatial attention together for better utilization of the pose information. Channel attention is mainly for features recalibration. In [Hu et al., 2018], the ‘squeeze-and-excitation’ block is proposed. It firstly aggregates the feature maps across the spatial dimensions and then passes them through a squeeze operation for producing the channel descriptor. The squeeze features are embeddings of the global distribution on the channel dimension. Then, an excitation operation, which takes the embeddings as input and produces a collection of channel modulation weights, is followed.

There are also works that make effort in utilizing both kinds of attention mechanisms. For instance: [Woo et al., 2018], both spatial and channel attention are introduced to enhance the performance of ResNet [He et al., 2016] architectures. [Chen et al., 2019] generalizes attention masks from spatial attention and channel attention to attention masks with high-order statistics information. [Li et al., 2018] propose the harmonious attention module that simultaneously learns ‘hard’ region level and ‘soft’ pixel-level attentive features for multi-granular feature representation.

3 The Proposed Method

In this section, we first introduce the architecture of our proposed Contextual Mutual Boosting Network (CMBN) and then elaborate on each of its components. As illustrated in Figure 1, apart from the shared convolutional frontend, there are two branches with different learning targets. The Backbone Branch (BB) (Sec.3.1) aims to learn the ID-discriminative foreground features of pedestrians for identification and the Reversal Branch (RB) (Sec.3.2) extracts the ID-irrelevant background features for boosting the BB’s pedestrian spatial localization indirectly. Meanwhile, at the end of the Backbone Branch, we propose the Mask Generator (MG) (Sec.3.3) to generate the static and universal channel mask for recalibrating the final representations. During the inference stage, both the RB and MG will be discarded. We regard the BB’s features multiplied by the channel mask as the final representation.

3.1 The Backbone Network

In this paper, we adopt the ResNet50 [He et al., 2016], which is well-known to extract the spatial features, as the backbone network. It consists of 1 convolutional block and 4 residual convolutional blocks named Conv1 ∼ 5 respectively. Here, we make three modifications on the original network: (1) we replace the Global Average Pooling (GAP) by the Global Maximum Pooling (2) we discard the final category classification layer, which is for classification on the ImageNet dataset [Deng et al., 2009], and add another two fully connected layers for producing the identity prediction of the input images (3) we remove the last spatial down-sampling to increase the granularity of features. The combination of the backbone network and FC layers is the IDE baseline [Zheng et al., 2016].

![Figure 2: Illustration of the Reversal Block.](image)

3.2 Reversal Branch

We aim to achieve the spatial localization function by the mutual boosting of the multiple contextual information. To achieve the goal, we propose the Reversal Branch (RB), which is the combination of the Convolutional Blocks and the Gradient Reversal Layer (GRL) [Ganin and Lempitsky, 2014], for learning the ID-irrelevant background features. It is known that GRL successfully combines domain adaptation and deep feature learning within one training process (deep domain adaptation). By maximizing the loss of the domain classifier, the network can not identify the domain categories correctly and thus generates the domain-invariant features. Following the idea, we introduce the GRL for enforcing the Reversal Branch to extract the ID-irrelevant background features, which are not optimal for identification, by maximizing the ID classification loss. More formally, the function can be formulated as:

\[
E(\theta_{id}, f) = -\lambda \sum_{i=1}^{N} L_y(G_{id}(f, \theta_{id}), y),
\]

where \(L_y\) is the Cross Entropy Loss, \(f\) is the feature representation of the images, \(y\) is the corresponding label, \(G_{id}\) is the ID classifier, and \(\theta_{id}\) is classifier parameters.

As shown in Figure 1, features extracted from the shared convolutional frontend \(Conv1 \sim 3\), will be sent into the BB and RB simultaneously. For the BB, features will go through the \(Conv4 \sim 5\). The obtained feature maps have two functions: (1) training the network’s pedestrian identification ability by minimizing the ID Loss (2) generating the reversal mask, whose spatial area roughly covers the ID-irrelevant background regions, for constraining the RB. To achieve the second function, we insert the Reversal Block, which is shown in Figure 2, between the BB and the RB. For the Reversal Block, it first adopts a learnable convolutional layer with \(1 \times 1\) kernel size for aggregating features along the channel dimension. A Batch Normalization (BN) is followed. Next, we apply the non-linear function \(Sigmoid\) for normalizing the value between 0 and 1.
the Reversal Operation by subtracting the value-limited output from 1 for getting the reversal mask. The function can be formulated as:

$$h = 1 - Sigmoid(Conv(fm)),$$

where $fm$ represents the feature maps from the BB, $Conv$ is the convolutional layer, $h$ is the generated mask. After getting the mask, an interpolation with a scale factor as 2 is adopted for enlarging the spatial size. The enlarged and original maps will be multiplied on the inputs of the $Conv_{4, Re}$ and $Conv_{5, Re}$ respectively for enforcing the spatial constrains. After going through all the $Conv_{Re}$ blocks, the GRL with hyperparameter $\lambda$ as 0.2 is inserted. Then, two FC layers for identity prediction is subsequent. During the forward propagation, the BB helps to shape the background spatial region of RB through the reversal masks. During the backpropagation, in verse, the RB helps to shape the foreground spatial region of BB by propagating the gradients through the Reversal Block. In this way, we achieve a mutual boosting scheme. The ID Loss will sustain a high value during the training process since the extracted features, which focus on the background information, lack the ID-relevant characteristics.

3.3 Mask Generator

Recalibration plays an important role in feature refinement. Previous works, including image classification [Woo et al., 2018; Hu et al., 2018], person Re-ID [Li et al., 2018; Chen et al., 2019] have widely adopted the attention mechanism for features recalibration. As discussed previously, many stacked SEblocks are required for noticeable improvement of the feature representation. To this end, we proposed a novel Mask Generator (MG), which aims to learn a static and universal mask for all the representations directly.

As we know, person Re-ID is a task that combines classification and retrieval together, meaning that no overlapping identities between training and testing set. We optimize the network by an additional FC layer, parametrized by a matrix $W$, for ID classification. In the training phase, the FC layer fits the data well and achieves 100% accuracy quickly. From the perspective of statistical inference, the activation distribution of the transformation matrix reflects the value of each channel indirectly. Our MG dedicates to exploring the statistical information for feature recalibration.

As shown in Figure 1, we achieve the function by adding a FC layer and a mask generator block. Specifically, we first adopt an additional FC layer, parameterized by a weight matrix, $W^{ID}$ for ID classification. The weight matrix and former network will be optimized by an ID Loss1. Then, we input the matrix $W^{ID}$ into the Mask Generator Block for generating the final mask. The block is constructed by another FC layer, parametrized by the weight matrix, $W^G$, and a $Sigmoid$ function. The matrix $W^G$ will generate a weight for each channel after inferring all the training identities’ feature activation distribution and $Sigmoid$ is for normalizing the weight to 0 and 1. The function can be formulated as:

$$Mask = Sigmoid(W^G(W^{ID})).$$

Since the generated mask will be multiplied on the representations, the MG will be optimized by the ID Loss2. The generated mask is not dynamically varying with the input, hence at the end of the training, there is no need to reserve the MG, we can only preserve the channel mask for recalibration through multiplication.

**Algorithm 1 Contextual-Detachment Strategy.**

**Input:**
- A mini-batch sample images of arbitrary identities; 

**Output:**
- CMBN; 

1: while not converge do; 
2: Optimizing the whole network with the two ID Losses and one ID Chaos Loss: 
   $$L = L_{ID1} + L_{ID2} + \beta L_{CS}$$
3: Fine-tuning the shared convolutional frontend and Backbone Branch with the two ID losses: 
   $$L = L_{ID1} + L_{ID2}$$
4: end while; 

3.4 Contextual-Detachment Strategy

As illustrated in Figure 1, two ID Losses ($L_{ID1}$ and $L_{ID2}$) and one ID Chaos Loss ($L_{CS}$) are used for optimizing the whole network. These three losses are all Cross Entropy Losses for classification. Note that the first two ID losses are applied to learn the foreground features, while the third one enables the Reversal Branch to focus on the ID-irrelevant background information.

![Figure 3: Visualization of the Backbone Branch’s activation maps.](image)

A straight forward idea is to jointly train the network with the three losses. To balance the conflicts of the two branches, a hyperparameter $\beta$ is multiplied on the $L_{CS}$. Here, we set $\beta = 0.4$. In Figure 3, we visualize the activation maps of the Backbone Branch’s outputs, which is optimized through the training strategy. We find that the background and occlusion pollute the feature maps significantly and some valuable characteristics are not salient enough. We argue that the problems are caused by the disharmonious cooperation between the two branches. Therefore, we propose the Contextual-Detachment Strategy (CDS), which is shown in Alg.1, to optimize the network. Specifically, we first jointly update the two branches with the three losses, which allows the two branches to focus on their corresponding contextual information and generate coarse activation maps. Then, we adopt only the two ID Losses to fine-tune the Backbone Branch. Through the training strategy, the BB can generate better-refined foreground features. In reverse, better foreground spatial localization boosts the background characteristics learning through the Reversal Block. The alternately optimization process is conducted end-to-end.
Figure 4: Visualization of the activation maps of the IDE and CMBN respectively. For better visual effects, we also map the activation maps on the original images.

| Methods | R=1 | mAP |
|---------|-----|-----|
| BoW+kissme [Zheng et al., 2015] | 44.4 | 20.8 |
| LOMO+XQDA [Liao et al., 2015] | 43.8 | 22.2 |
| MGCAM [Song et al., 2018] | 83.79 | 74.33 |
| SPReID [Kalayeh et al., 2018a] | 93.68 | 83.36 |
| PDC [Su et al., 2017] | 84.14 | 63.41 |
| AACN [Xu et al., 2018] | 85.90 | 66.87 |
| PN-GAN [Qian et al., 2018] | 89.43 | 72.58 |
| PGFA [Miao et al., 2019] | 91.2 | 76.8 |
| HA-CNN [Li et al., 2018] | 91.2 | 75.7 |
| DuATM [Si et al., 2018] | 91.4 | 76.6 |
| CASN (IDE) [Zheng et al., 2019] | 92.0 | 78.0 |
| CMBN | 92.8 | 79.8 |

| Methods | R=1 | mAP |
|---------|-----|-----|
| BoW+kissme [Zheng et al., 2015] | 25.1 | 12.2 |
| LOMO+XQDA [Liao et al., 2015] | 30.8 | 17.0 |
| SPReID [Kalayeh et al., 2018a] | 85.95 | 73.34 |
| PN-GAN [Qian et al., 2018] | 73.58 | 53.20 |
| AACN [Xu et al., 2018] | 76.84 | 59.25 |
| PGFA [Miao et al., 2019] | 82.6 | 65.5 |
| HA-CNN [Li et al., 2018] | 80.5 | 63.8 |
| DuATM [Si et al., 2018] | 81.8 | 64.6 |
| CASN (IDE) [Zheng et al., 2019] | 84.5 | 67.0 |
| CMBN | 84.8 | 69.5 |

| Methods | Detected R=1 | mAP | Labeled R=1 | mAP |
|---------|---------------|-----|-------------|-----|
| BoW+XQDA [Zheng et al., 2015] | 6.4 | 6.4 | 7.9 | 7.3 |
| LOMO+XQDA [Liao et al., 2015] | 12.8 | 11.5 | 14.8 | 13.6 |
| MGCAM [Song et al., 2018] | 46.71 | 46.87 | 50.14 | 50.21 |
| HA-CNN [Li et al., 2018] | 41.7 | 38.6 | 44.4 | 41.0 |
| CASN (IDE) [Zheng et al., 2019] | 57.4 | 50.7 | 58.9 | 52.2 |
| CMBN | 63.1 | 58.9 | 67.5 | 63.0 |

4 Experiments

4.1 Datasets and Evaluation Measures

Datasets. In this paper, we adopt the Market-1501 [Zheng et al., 2015], CUHK03-NP [Li et al., 2014] and DukeMTMC-ReID [Gou et al., 2017] datasets for experiments. Market-1501 collects totally 12,936 training images across 6 cameras without overlapping views. As for the testing data, gallery and query sets consist of 19,732 and 3,368 images respectively with 750 different identities. CUHK03-NP adopts a new split protocol for the CUHK03. Following the new protocol, there are 767/700 identities for training and testing respectively. DukeMTMC-ReID includes 16,522 training images of 702 identities, 2,228 query and 17,661 gallery images of another 702 identities.

Evaluation Measures. Cumulative matching characteristics (CMC) cure and mAP are adopted for evaluation. They present the accuracy of the retrieval and the recall rate, respectively.

4.2 Implementation Details

The proposed Contextual Mutual Boosting Network (CMBN) is based on the ResNet50 architecture. We fine-tune them with the AMSGrad on the Re-ID datasets. The learning rate is initialized as 0.0003 and decayed by 0.1 every 10 epochs.

4.3 Visualization of the Activation Maps

In Figure 4, we visualize the activation maps from the Backbone Branch. For each example, the activation maps of the IDE and our CMBN are illustrated respectively. The IDE has two obvious shortcomings: (1) the salient features are not...
consistent, many potentially informative cues are ignored, (2) distracting factors like clutting background pollute the representations enormously. Our proposed CMBN, especially the Reversal Branch and the Contextual-Detachment Strategy, effectively ameliorates the aforementioned two problems.

4.4 Comparison with the state-of-the-art

To demonstrate the effectiveness of the proposed Contextual Mutual Boosting Network (CMBN), we compare our architecture with the state-of-the-art methods on the Market-1501, DukeMTMC-ReID, and CUHK03-NP datasets and the results are present in Table 1, 2, 3, respectively. Noted that, we compare methods based on the original classification network like ResNet rather than other superior modified networks like Long et al.

| Methods                  | Market1501 Rank-1 | Market1501 mAP | DukeMTMC-ReID Rank-1 | DukeMTMC-ReID mAP | CUHK03-NP Detected Rank-1 | CUHK03-NP Detected mAP | CUHK03-NP Labeled Rank-1 | CUHK03-NP Labeled mAP |
|--------------------------|------------------|----------------|----------------------|-------------------|--------------------------|-------------------------|-------------------------|------------------------|
| BaseLine                 | 89.5             | 70.7           | 80.2                 | 62.3              | 50.8                     | 46.0                    | 53.1                    | 48.1                   |
| BaseLine+RB              | 91.4             | 77.8           | 84.1                 | 67.5              | 59.4                     | 55.1                    | 60.4                    | 56.9                   |
| BaseLine+MG              | 91.7             | 78.9           | 84.8                 | 69.0              | 60.9                     | 57.4                    | 64.8                    | 61.7                   |
| BaseLine+RB+MG           | 91.9             | 78.8           | 84.3                 | 68.4              | 60.1                     | 56.2                    | 63.3                    | 60.2                   |
| BaseLine+RB+MG+CDS       | **92.8**         | **79.8**       | **84.8**             | **69.5**          | **63.1**                 | **58.9**                | **67.5**                | **63.0**               |

4.5 Component Analysis

In this section, we further verify the effectiveness of the proposed ‘Reversal Branch (RB)’, ‘Mask Generator (MG)’, and the ‘Contextual-Detachment Strategy (CDS)’ respectively. In Table 4, we report the evaluation results of the proposed model on CUHK03-NP, Market-1501, and DukeMTMC-ReID, and work up to the final CMBN step by step. Here, We adopt the modified ResNet50 with two additional FC layers (IDE) as the ‘BaseLine’. From Table 4, we can see clear improvements in performance over the ‘BaseLine’ with the proposed models.

For instance on the CUHK03-NP (Labeled) dataset, after adding the ‘RB’, the Rank-1 and mAP improves by 7.3% and 8.8%, respectively. This provides that localizing pedestrians through mutual contextual boosting does improve the discrimination of representations. It’s an important and integral part of our proposed CMBN. Furthermore, we also test adding the ‘MG’ and there are also 11.7% and 13.6% improvements of Rank-1 and mAP, respectively. It demonstrates that the channel mask is effective in recalibrating the obtained features. However, after combing the ‘RB’ and ‘MG’ together, the accuracies do not increase but drop compared with the ‘BaseLine+MG’. It means that directly combing the ‘RB’ and ‘MG’ together will pollute the representations since the noise is introduced into the BB from the shared frontend. Therefore, it is necessary to enforce the foreground learning constraints to the BB especially. After adding the ‘CDS’, the Rank-1 and mAP both reach the highest points, proving the importance of the alternately training strategy.

5 Conclusion

In this paper, we propose the Contextual Mutual Boosting Network, which aims to generate discriminative features of pedestrians for identification. We construct the Reversal Branch, which aims to learn the background information, into the network. Through the mutual contextual boosting, the network localizes the spatial regions of pedestrians more accurately. Besides, we propose a unique Mask Generator for feature recalibration on the channel dimension by exploiting the statistical inference. At last, we propose the Contextual-Detachment Strategy for optimizing the network. It optimizes the network by learning the multiply contextual information first and then refining the foreground features only. In the inference phase, the Reversal Branch and Mask Generator will be removed, only the shared convolutional frontend, the Backbone Branch, and the generated channel mask...
are needed, which makes our network more computationally efficient and practical. Extensive experiments on the benchmark Re-ID datasets demonstrate the competitive performance compared with other state-of-the-art methods.

References

[Chen et al., 2019] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In ICCV, pages 371–381, 2019.

[Cho and Yoon, 2016] Yeong-Jun Cho and Kuk-Jin Yoon. Improving person re-identification via pose-aware multi-shot matching. In CVPR, pages 1354–1362, 2016.

[Deng et al., 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, pages 248–255, 2009.

[Ganin and Lempitsky, 2014] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495, 2014.

[Gou et al., 2017] Mengran Gou, Srikrishna Karanam, Wen-qian Liu, Octavia Camps, and Richard J Radke. Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. In CVPR Workshops, pages 10–19, 2017.

[He et al., 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016.

[Hu et al., 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In CVPR, pages 7132–7141, 2018.

[Kalayeh et al., 2018a] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In CVPR, pages 1062–1071, 2018.

[Kalayeh et al., 2018b] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In CVPR, pages 1062–1071, 2018.

[Li et al., 2014] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In CVPR, 2014.

[Li et al., 2018] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In CVPR, pages 2285–2294, 2018.

[Liao et al., 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In CVPR, pages 2197–2206, 2015.

[Liu et al., 2018] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In CVPR, pages 4099–4108, 2018.

[Miao et al., 2019] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In ICCV, pages 542–551, 2019.

[Qi et al., 2018] Lei Qi, Jing Huo, Lei Wang, Yinghuan Shi, and Yang Gao. Maskreid: A mask based deep ranking neural network for person re-identification. arXiv preprint arXiv:1804.03864, 2018.

[Qian et al., 2018] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In ECCV, pages 650–667, 2018.

[Si et al., 2018] 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 CVPR, pages 5363–5372, 2018.

[Song et al., 2018] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In CVPR, pages 1179–1188, 2018.

[Su et al., 2017] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In ICCV, pages 3960–3969, 2017.

[Sun et al., 2018] 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 ECCV, pages 480–496, 2018.

[Wang et al., 2018a] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In ACM Multimedia, pages 274–282, 2018.

[Wang et al., 2018b] Haoran Wang, Yue Fan, Zexin Wang, Licheng Jiao, and Bernt Schiele. Parameter-free spatial attention network for person re-identification. arXiv preprint arXiv:1811.12150, 2018.

[Woo et al., 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In ECCV, pages 3–19, 2018.

[Xu et al., 2018] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In ECCV, 2018.

[Zeiler and Fergus, 2014] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In ECCV, pages 818–833, 2014.

[Zhao et al., 2017] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jin Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In CVPR, pages 1077–1085, 2017.

[Zheng et al., 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In CVPR, 2015.
[Zheng et al., 2016] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984, 2016.

[Zheng et al., 2019] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J Radke. Re-identification with consistent attentive siamese networks. In CVPR, pages 5735–5744, 2019.