Towards Good Practices on Building Effective CNN Baseline Model for Person Re-identification

FU XIONG¹, YANG XIAO¹,*, ZHIGUO CAO¹, KAICHENG GONG¹, ZHIWEN FANG², AND JOEY TIANYI ZHOU³

¹ National Key Laboratory of Science and Technology on Multi-Spectral Information Processing, School of Automation, Huazhong University of Science and Technology
² School of Energy and Mechanical-electronic Engineering, Hunan University of Humanities, Science and Technology, Loudi, China
³ Institute of High Performance Computing, A*STAR, Singapore

Corresponding author: Yang Xiao@hust.edu.cn

1. INTRODUCTION

As a fine-grained visual recognition problem, person re-identification is indeed a challenging visual recognition task due to the critical issues of human pose variation, human body occlusion, camera view variation, etc. To address this, most of the state-of-the-art approaches are proposed based on deep convolutional neural network (CNN), being leveraged by its strong feature learning power and classification boundary fitting capacity. Although the vital role towards person re-identification, how to build effective CNN baseline model has not been well studied yet. To answer this open question, we propose 3 good practices in this paper from the perspectives of adjusting CNN architecture and training procedure. In particular, they are adding batch normalization after the global pooling layer, executing identity categorization directly using only one fully-connected, and using Adam as optimizer.

The extensive experiments on 3 widely-used benchmark datasets demonstrate that, our propositions essentially facilitate the CNN baseline model to achieve the state-of-the-art performance without any other high-level domain knowledge or low-level technical trick.

1. INTRODUCTION

As a fine-grained visual recognition problem, person re-identification is indeed a challenging visual recognition task due to the critical issues of human pose variation, human body occlusion, camera view variation, etc. To address this, most of the state-of-the-art approaches are proposed based on deep convolutional neural network (CNN), being leveraged by its strong feature learning power and classification boundary fitting capacity. Although the vital role towards person re-identification, how to build effective CNN baseline model has not been well studied yet. To answer this open question, we propose 3 good practices in this paper from the perspectives of adjusting CNN architecture and training procedure. In particular, they are adding batch normalization after the global pooling layer, executing identity categorization directly using only one fully-connected, and using Adam as optimizer.

The extensive experiments on 3 widely-used benchmark datasets demonstrate that, our propositions essentially facilitate the CNN baseline model to achieve the state-of-the-art performance without any other high-level domain knowledge or low-level technical trick.

Person re-identification is indeed a challenging visual recognition task due to the critical issues of human pose variation, human body occlusion, camera view variation, etc. To address this, most of the state-of-the-art approaches are proposed based on deep convolutional neural network (CNN), being leveraged by its strong feature learning power and classification boundary fitting capacity. Although the vital role towards person re-identification, how to build effective CNN baseline model has not been well studied yet. To answer this open question, we propose 3 good practices in this paper from the perspectives of adjusting CNN architecture and training procedure. In particular, they are adding batch normalization after the global pooling layer, executing identity categorization directly using only one fully-connected, and using Adam as optimizer.

The extensive experiments on 3 widely-used benchmark datasets demonstrate that, our propositions essentially facilitate the CNN baseline model to achieve the state-of-the-art performance without any other high-level domain knowledge or low-level technical trick.

© 2021 Optical Society of America

http://dx.doi.org/10.1364/ao.XX.XXXXXX
2. RELATED WORKS

In this section we mainly focus on deep learning baseline for person re-identification, and existing Approaches for preventing over fitting. [2] is recommend if the readers are interested in an overall review of re-ID.

A. Baseline for deep person re-identification

Traditionally, hand-crafted feature designed based on color histogram prevails in person re-ID [1, 13, 14], due to the observation that color of clothes has good discriminability for distinguishing persons. Recent researches on person re-ID mostly focus on building deep convolutional neural network in the end-to-end learning manner. [2] takes advantage of the deep convolutional models pre-trained on ImageNet [5] and fine-tunes it on person re-identification datasets using softmax loss. The feature from the last pooling layer is used as the image representation. The learned representation achieves great performance boost against traditional hand-crafted feature. Due to the success of [2], most current deep learning based methods also adopt pre-trained models as backbone network and have been searching other technical means to further leverage the performance of re-ID system. Hence, for most current approaches [3, 15, 16] feature learned using only softmax loss usually serves as a baseline for comparison.

Different network architectures have been explored in person re-identification filed. Among existing person re-identification approaches, Resnet50 [10] is the most commonly used backbone network [15, 16]. Besides, GoogLeNet [17], inception networks [18] and densenet [12] have also been chosen as backbone network by some researchers. Taking advantage of the pre-trained CNN models, by further employing metric learning methods [19], using part-based CNN representation[20] or carefully designing attention mechanism[21], the performance for person re-identification can be further improved.

B. Approaches for preventing overfitting

To prevent overfitting for deep CNN models when trained on relatively small datasets, many approaches have been proposed. In particular, random cropping [22], random flipping [23] and random erase operation [24] are commonly used data augmentation methods in training deep CNN model. Besides, regularization methods like weight decay is also a well-known approach for prevent overfitting. Recently, batch normalization and dropout are two widely used tricks for training CNN, and have show benefits for preventing overfitting. Dropout randomly discards the output of each hidden neuron with a probability during the training process. Batch normalization aims at reducing internal co-variate shift by normalizing the output of each hidden neuron using mini-batch mean and variance. Since person re-identification dataset are relatively small (e.g Market1501 containing only 12,936 images for training), effective means for preventing overfitting is necessary for building high-accuracy person re-identification model.

3. GOOD PRACTICES ON BUILDING EFFECTIVE CNN BASELINE MODEL

In this section, we will illustrate 3 key practices towards building effective CNN baseline model for person re-identification. The main technical pipeline of out approach and our proposed 3 good practices are shown in Fig. 1. First, the inputs are fed into the backbone CNN network (e.g., Resnet50). Then, the “global pooling feature” yielded by the last pooling layer is fed into the batch normalization layer to generate the final subject representation. With the batch-normalized feature, we directly execute person identity categorization using only one fully-connected layer. That is, the procedure of dimension reduction using multiple fully-connected layers [15, 25] is removed. At last, Adam is employed as the optimizer to train CNN model. Intuitively, our proposition is simple but can be easily applied to various kinds of CNN architectures. It is worthy noting that, only softmax loss is used as supervision signal in our approach.

Although the 3 practices are proposed mainly according to our experience and experiments, we guess the superiority of these 3 practices lies on preventing overfitting and maintaining the discriminative power of the model pre-trained on ImageNet. ImageNet dataset contains over 14 million images, which has provided plenty of visual concepts. Recently, models pre-trained on ImageNet have demonstrated good transfer ability in various computer vision tasks [26, 27]. Hence, we think a carefully designed fine-tuning procedure can also benefit person re-ID by preserving the discriminative power of pre-trained model and preventing overfitting. Next, we will illustrate the proposed 3
good practices in details.

A. Execute batch normalization after global feature pooling

To well fine-tune CNN model pre-trained on ImageNet towards the relative small-scale datasets (e.g., person re-identification dataset), one critical issue is to alleviate overfitting problem during training. Current dropout [28] and batch normalization [29] are two widely used techniques to address this. Dropout randomly discards the output of the hidden neuron with a probability during the training process. Batch normalization aims at reducing internal covariate shift by normalizing the output of each hidden neuron using mini-batch mean and variance.

For the specific person re-identification task, some works [30] adopt dropout after the global pooling layer in order to prevent overfitting. In our implementation, we empirically choose to execute batch normalization after the global feature pooling layer instead of dropout, as shown in Fig. 1. In the testing phase, the feature after the batch-normalization layer is chosen as the person image representation. Compared with dropout, which needs to randomly discards some hidden neuron, batch normalization can provide more steady gradient. This can alleviate the unnecessary disturbance to the pre-trained model.

In our experiments we find that, batch normalization generally leads to faster convergence speed and better performance. The experiments conducted in Sec. 4 will verify the effectiveness of our proposition.

B. Conduct identity categorization directly using only one fully-connected layer

Within some well-established person re-identification CNN models [15], 2 fully-connected layers are usually set after the global feature pooling layer. The first fully-connected layer plays the role of “bottleneck” to conduct feature dimension reduction. And, the second fully-connected layer executes person identity categorization. However, according to our experience, the introduction of “bottleneck” fully-connected layer often decreases the final performance essentially. Thus, we choose to remove the “bottleneck” layer. With the batch normalized global pooling feature yielded in Sec. A, we propose to conduct person identity categorization directly using only one fully-connected layer.

It is worth noting that, our proposition actually helps to compress the employed CNN model. By discarding the “bottleneck” layer, the gradient from the softmax loss can be directly pass backward to the convolutional layer. This also essentially benefits to alleviate overfitting problem to ensure the test performance.

C. Optimize CNN model using Adam

The last key practice is to train CNN model using Adam as the optimizer, under the supervision of cross entropy loss. Adam is the recently proposed first-order gradient-based optimization method for stochastic objective function. It executes based on the adaptive estimates of lower-order moments. In particular, Adam update the parameters according to Eq. 1.

$$
\begin{align*}
& t \leftarrow t + 1 \\
& g_t \leftarrow \nabla_{\theta} f_{\theta}(\theta_{t-1}) \\
& m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \\
& v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \\
& \hat{m}_t \leftarrow m_t / (1 - \beta_1^t) \\
& \hat{v}_t \leftarrow v_t / (1 - \beta_2^t) \\
& \theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon),
\end{align*}
$$

where $t$ denotes the timestep. $f$ means the objective function. $\theta$ represents the learnable parameters. $\alpha$ denotes the learning rate. $\beta_1$, $\beta_2$ and $\epsilon$ are hyper-parameters, whose default setting is $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e - 8$.

Currently, most of the state-of-the-art person re-identification methods [4, 15] choose stochastic gradient descent (SGD) to train CNN model. Compared with SGD, Adam adaptive estimates lower-order moments, which smooths the variation of the gradients. This can also alleviate the unnecessary disturbance to the pre-trained model as well as prevent over-fitting. The experiments conducted in Sec. 4 will verify the effectiveness of our proposition.

4. EXPERIMENTS

A. Experimental setup

Datasets: To evaluate the effectiveness of our propositions, we conduct experiments on three widely used person re-identification datasets, including Market-1501 [7], DukeMTMC-reID [8] and CUHK03 [9]. The Market-1501 dataset consists of image samples from 6 cameras with different resolutions. It contains 1,501 identities and 32,688 bounding boxes. The training set contains 12,936 bounding boxes of 751 identities. The remaining 750 identities are used for test. DukeMTMC-reID dataset contains 1,812 identities with 8 cameras. There are 1,404 identities that appear in more than two cameras, and 408 identities (i.e., distractor ID) that appear in only one camera. 702 identities are selected for training, and the remaining 702 identities as used for test. CUHK03 dataset consists of 14,097 image samples captured by 6 cameras from 1,467 persons. Two kinds of annotations are provided in this dataset: manually labeled pedestrian bounding boxes and DPM-based [31] bounding boxes. On CUHK03 dataset, we use the test protocol in [9]. That is, image samples from 767 identities are selected for training. The other 700 identities are employed for test. Fig. 2 shows the image samples from these 3 datasets. It can be seen that, the issues of human pose variation, human body occlusion and camera view variation impose great challenge to this task.

Evaluation: In testing phase, Euclidean distance of $L_2$ normalized features is used as similarity metric. Cumulated Matching Characteristics (CMC) curve is employed to evaluate the performance of person re-identification methods. The CMC curve shows the probability that a query identity appears in different-sized candidate lists. Because of the space limitation, we only report the cumulated matching accuracy at selected ranks rather than plotting the actual curves. Mean average precision is also used as a performance measure. Mean average precision indicates the area under the Precision-Recall curve.

Implementation details: Our CNN-based person re-identification approach is implemented using PyTorch framework. The CNN models used by us are pre-trained on ImageNet.
The input image is resized to 256×128 and padded to 276×148 with zero. Input images are then randomly left-right flipped and cropped to 256×128 for data augmentation. Left-right image flipping is also used in testing phase. Training iterations are set as 60 epochs for all datasets. The Adam [6] optimiser is used with a mini-batch size of 32. Following [3], the initial learning rate is set as 0.00035 and weight decay is set as 5e-4 for all datasets. Adam parameters are set as $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The decay factor for the learning rate is set as 0.1 at every 20 epochs.

**B. Results on Market-1501 dataset**

Performance comparison between our method and the other state-of-the-art approaches on Market-1501 dataset is shown in Table 1. Specifically, R50 represents Resnet50 network [10], R39 represents Resnet39 network [10], Rx50 represents Resnext50 network [11], Dse121 represents Densenet121 network [12], GN represents GoogLeNet network [17], and Icp-V3 represents Inception-V3 network [18]. We do not report the CNN architecture if the corresponding method does not use pre-trained model. To verify the effectiveness and generality of our proposition, it is applied to multiple CNN architectures (i.e., Resnet50, Resnext50 and Densenet121) widely used for person re-identification task. It can be observed that, being facilitated by the proposed good practices Resnet50, Resnext50 and Densenet121 outperform all the state-of-the-art approaches in all the test cases.

It is worthy noting that, DaRe [32], CamStyle+RE [33, 34], MLFN [3], HA-CNN [21] and DuATM [4] are the most recently published approaches at CVPR 2018. In particular, DaRe fuses the feature from the multiple layers of Resnet50. CamStyle+RE takes advantage of cycle-gan [35] to generate samples of different camera styles, and then uses them to fine-tuning CNN in the idea of data augment. MLFN, HA-CNN and DuATM design extra attention mechanism to acquire discriminative feature. However, we train CNN model only use softmax loss. No metric learning methods, no global local fusion and no attention mechanism are applied.

**Table 1. Performance comparison on Market-1501 dataset.**

| Method | Single-query | Multi-query |
|--------|--------------|-------------|
|        | Rank-1 | mAP | Rank-1 | mAP |
| LOMO+XQDA [1] | 43.8 | 22.2 | 54.1 | 28.4 |
| IDE [2] | 73.9 | 47.8 | – | – |
| Spindle [36] | 76.9 | – | – | – |
| OIM [37] | 82.1 | – | – | – |
| Re-rank [9] | 77.1 | 63.6 | – | – |
| DPA [20] | 81.0 | 63.4 | – | – |
| SVDNet [15] | 82.3 | 62.1 | – | – |
| IDE+DaF [16] | 82.3 | 72.4 | – | – |
| ACRN [38] | 83.6 | 62.6 | – | – |
| Context [39] | 80.3 | 57.5 | 86.8 | 66.7 |
| JLML [40] | 83.9 | 64.4 | 89.7 | 74.5 |
| SSM [41] | 82.2 | 68.8 | 88.2 | 76.2 |
| DaRe [32] | 86.4 | 69.3 | – | – |
| DPFL [42] | 88.6 | 72.6 | 92.2 | 80.4 |
| CamStyle+RE [33, 34] | 89.5 | 71.6 | – | – |
| MLFN [3] | 90.0 | 74.3 | 92.3 | 82.4 |
| HA-CNN [21] | 91.2 | 75.7 | 93.8 | 82.7 |
| DuATM [4] | 91.4 | 76.6 | – | – |

**Table 2. Performance comparison on DukeMTMC-reID dataset.**

| Methods | Rank-1 | mAP | Backbone |
|---------|--------|-----|----------|
| BoW+kissme [7] | 25.1 | 12.2 | – |
| LOMO+XQDA [1] | 30.8 | 17.0 | – |
| IDE [2] | 65.2 | 45.0 | R50 |
| ACRN [38] | 72.6 | 52.0 | GN |
| SVDNet [15] | 76.7 | 56.8 | R50 |
| AACN [43] | 76.8 | 59.3 | GN |
| CamStyle+RE [33, 34] | 78.3 | 57.6 | R50 |
| DPFL [42] | 79.2 | 60.6 | Icp-V3 |
| PSE [44] | 79.8 | 62.0 | R50 |
| ATWL [45] | 79.8 | 63.4 | R50 |
| DaRe [32] | 75.2 | 57.4 | R50 |
| HA-CNN [21] | 80.5 | 63.8 | – |
| MLFN [3] | 81.2 | 62.8 | Rs50 |
| DuATM [4] | 81.8 | 64.6 | Dse121 |

| Resnet50 (ours) | 83.4 | 68.8 | R50 |
| Resnext50 (ours) | 82.8 | 68.1 | Rs50 |
| Densenet121 (ours) | 83.5 | 68.5 | Dse121 |
According to our observation, we categorize the failure cases into four categories, they are: (1) failure cases caused by occlusion; (2) failure cases caused by identities with similar-looking people; (3) failure cases caused by multi-person query images. (4) inexpllicable failure cases. next we will show them respectively. To quantify the four cases, We have also counted the percentage of each case, which is shown in Table 7.

Failure cases 1 – caused by occlusion: When the person in the image, especially the query image, is heavily occluded, the

Table 3. Performance Comparison on CUHK03 dataset.

| Method          | labeled rank-1 | mAP | detected rank-1 | mAP | Backbone |
|-----------------|----------------|-----|-----------------|-----|----------|
| BOW+XQDA [7]    | 7.9            | 7.3 | 6.4             | 6.4 | –        |
| LOMO+XQDA [1]   | 14.8           | 13.6| 12.8            | 11.5| –        |
| IDE [2]         | 22.2           | 21.0| 21.3            | 19.7| R50      |
| IDE+DaF [16]    | 27.5           | 31.5| 26.4            | 30.0| R50      |
| DFPI [42]       | 43.0           | 40.5| 40.7            | 37.0| Icp-V3   |
| SVDNet [15]     | 40.9           | 37.8| 41.5            | 37.3| R50      |
| HA-CNN [21]     | 44.4           | 41.0| 41.7            | 38.6| –        |
| MLIN [3]        | 54.7           | 49.2| 52.8            | 47.8| Rx50     |
| Duke [32]       | 58.1           | 53.7| 55.1            | 51.3| R50      |
| Resnet50(ours)  | 62.1           | 58.1| 56.3            | 52.2| R50      |
| Resnext50(ours) | 60.1           | 56.2| 59.1            | 54.4| Rx50     |
| Densenet121(ours)| 63.5          | 59.0| 57.8            | 52.8| Dse121   |

Table 4. Ablation study on practice 1.

| Dataset          | Market-1501 rank-1 | mAP | DukeMTMC-reID rank-1 | mAP |
|------------------|---------------------|-----|----------------------|-----|
| w/o BN           | 78.6                | 57.9| 70.0                 | 50.9|
| Dropout          | 79.2                | 59.9| 72.9                 | 55.4|
| Good practices   | 91.7                | 78.8| 83.4                 | 68.8|

Table 5. Ablation study on practice 2.

| Dataset          | Market-1501 rank-1 | mAP | DukeMTMC-reID rank-1 | mAP |
|------------------|---------------------|-----|----------------------|-----|
| Bottleneck       | 89.4                | 73.4| 78.6                 | 61.8|
| Good practices   | 91.7                | 78.8| 83.4                 | 68.8|

Table 6. Ablation study on practice 3.

| Dataset          | Market-1501 rank-1 | mAP | DukeMTMC-reID rank-1 | mAP |
|------------------|---------------------|-----|----------------------|-----|
| SGD              | 88.9                | 72.4| 80.0                 | 64.5|
| Good practices   | 91.7                | 78.8| 83.4                 | 68.8|

Table 7. Percentage of each category of the failure cases.

| Fail cases | category 1 | category 2 | category 3 | category 4 |
|------------|------------|------------|------------|------------|
| Percentage | 19.6%      | 34.3%      | 37.8%      | 8.3%       |
Occlusion affects the performance in two aspects. First, occlusion makes the description of the person not reliable, because the useful information only comes from the visible part of the body. Secondly, since the representation is obtained using global pooling, the feature of the occlusion objects may also be encoded into the representation. This leads to inaccurate description of the person. This category of failure cases account for 19.6% of all failure cases.

Failure cases 2 – caused by similar looking: How to identify persons with similar looking is an inherent difficulty in person re-identification. Actually, our model achieves rank-1 accuracy of 83.4%. So the failure cases illustrated in fig 4 will tell us what kind of images pairs are still challenging for the state-of-the-art person re-identification model. This category of failure cases account for 34.3% of all failure cases, which we think is the major challenge of person re-identification.

Failure cases 3 – multiple persons in one query images: The third category of failure cases arise from multiple person in one query image, which is illustrated in Fig 5. We think this kind of failure is tolerable though it account most for the failure cases, because in this case it is ambiguous that which person is the one for query. And in many failure cases, in spite that the searched person are not the person for query, it is indeed in the query image.

Failure cases 4 – inexplicable failure cases: The failure cases of this category are illustrated in Fig 6. It is hard to name the cause of this category of failure cases. These failure cases are not hard to discriminate for human beings, but the CNN model fails to identify them. Hence, we call it inexplicable failure cases. Good news is that this category of failure cases account the least, only 8.3%, for of all failure cases.

5. CONCLUSIONS
In this paper, we have proposed 3 good practices for building effective CNN baseline model towards person re-identification, including adding batch normalization layer after the global pooling layer, conducting identity categorization directly using only one fully-connected layer and using Adam as optimizer. By applying these 3 practices, we have achieved state-of-the-art performance on 3 challenging datasets. We believe that our proposition can serve as a good baseline to promote the research on person re-identification in future.

ACKNOWLEDGMENT
This work is jointly supported by the National Key R&D Program of China (No. 2018YFB1004600), National Natural Science Foundation of China (Grant No. 61502187, 61602193 and 61702182), the International Science & Technology Cooperation Program of Hubei Province, China (Grant No. 2017AHB051), the HUST Interdisciplinary Innovation Team Foundation (Grant No. 2016JCTD120), Hunan Provincial Natural Science Foundation of China (Grant 2018JJ3254).

Fig. 3. Failure cases 1: failure cases caused by occlusion.

Fig. 4. Failure case 2: different person with similar looking.

Fig. 5. Failure case 3: multiple persons in one query image.

Fig. 6. Failure case 4: inexplicable failure cases.
REFERENCES

1. S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2015), pp. 2197–2206.

2. L. Zheng, Y. Yang, and A. G. Hauptmann, arXiv preprint arXiv:1610.02984 (2016).

3. X. Chang, T. M. Hospedales, and T. Xiang, “Multi-level factorisation net for person re-identification,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1 (2018), p. 2.

4. J. Shi, 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,” (2018).

5. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (IEEE, 2009), pp. 248–255.

6. D. P. Kingma and J. Ba, arXiv preprint arXiv:1412.6980 (2014).

7. L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in Proc. IEEE International Conference on Computer Vision (ICCV), (2015), pp. 1116–1124.

8. Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by gan improve the person re-identification baseline in vitro,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2017), pp. 3754–3762.

9. Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with k-reciprocal encoding,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (IEEE, 2017), pp. 3652–3661.

10. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016), pp. 770–778.

11. S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (IEEE, 2017), pp. 5987–5995.

12. G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, “Densely connected convolutional networks,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1 (2017), p. 3.

13. F. Jurie and A. Mignon, “Pcca: A new approach for distance learning from sparse pairwise constraints,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition, (2012), pp. 2666–2673.

14. D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” in Proc. IEEE Conference on European Conference on Computer Vision, (2008), pp. 262–275.

15. Y. Sun, L. Zheng, W. Deng, and S. Wang, “Sydnet for pedestrian retrieval,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2017), pp. 3800–3808.

16. R. Yu, Z. Zhou, S. Bai, and X. Bai, arXiv preprint arXiv:1708.04169 (2017).

17. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich et al., “Going deeper with convolutions,” (Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2015).

18. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proc. IEEE conference on Computer Vision and Pattern Recognition(CVPR), (2016), pp. 2818–2826.

19. A. Herrmans, L. Beyer, and B. Leibe, arXiv preprint arXiv:1703.07737 (2017).

20. L. Zhao, X. Li, Y. Zhuang, and J. Wang, “Deeply-learned part-aligned representations for person re-identification,” in Proc. IEEE International Conference on Computer Vision (ICCV), (2017), pp. 3239–3248.

21. W. Li, X. Zhu, and S. Gong, “Harmonious attention network for person re-identification,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1 (2018), p. 2.

22. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. IEEE Conference on International Conference on Neural Information Processing Systems, (2012), pp. 1097–1105.

23. K. Simonyan and A. Zisserman, Comput. Sci. (2014).

24. Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, arXiv preprint arXiv:1708.04896 (2017).

25. K. Simonyan and A. Zisserman, arXiv preprint arXiv:1409.1556 (2014).

26. J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition, (2015), pp. 3431–3440.

27. J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, (2017), pp. 6517–6525.

28. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, J. Mach. Learn. Res. 15, 1929 (2014).

29. S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. International Conference on Machine Learning(IJCL), (2015), pp. 448–456.

30. Z. Zheng, L. Zheng, and Y. Yang, ACM Transactions on Multimodal. Comput. Commun. Appl. (TOMM) 14, 13 (2017).

31. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, IEEE Transactions on Pattern Analysis Mach. Intell. 32, 1627 (2010).

32. Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger, “Resource aware person re-identification across multiple resolutions,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), (2018), pp. 8042–8051.

33. Z. Zhong, L. Zheng, Z. Zhang, S. Li, and Y. Yang, arXiv preprint arXiv:1711.10295 (2017).

34. Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, arXiv preprint arXiv:1709.04896 (2017).

35. J. Yu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), (2017), pp. 2223–2232.

36. 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 Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), (2017), pp. 907–915.

37. T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, “Joint detection and identification feature learning for person search,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (IEEE, 2017), pp. 3376–3385.

38. A. Schumann and R. Stiefelhagen, “Person re-identification by deep learning attribute-complementary information,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW), (2017), pp. 1435–1443.

39. D. Li, X. Chen, Z. Zhang, and K. Huang, “Learning deep context-aware features over body and latent parts for person re-identification,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2017).

40. W. Li, X. Zhu, and S. Gong, “Person re-identification by deep joint learning of multi-loss classification,” in Twenty-Sixth International Joint Conference on Artificial Intelligence, (2017), pp. 2194–2200.

41. S. Bai, X. Bai, and Q. Tian, “Scalable person re-identification on supervised smoothed manifold,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), vol. 6 (2017), p. 10.

42. Y. Chen, X. Zhu, and S. Gong, “Person re-identification by deep learning multi-scale representations,” in Proc. IEEE International Conference on Computer Vision Workshops(ICCVW), (2018), pp. 2590–2600.

43. J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, arXiv preprint arXiv:1805.03344 (2018).

44. 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 Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), (2018), pp. 420–429.

45. E. Ristani and C. Tomasi, arXiv preprint arXiv:1803.10859 (2018).