Person Re-identification by Mid-level Attribute and Part-based Identity Learning

Guopeng Zhang 2447676153@qq.com

Jinhua Xu jhxu@cs.ecnu.edu.cn

Department of Computer Science and Technology
Shanghai Key Laboratory of Multidimensional Information Processing
East China Normal University, Shanghai 200062, China

Editors: Jun Zhu and Ichiro Takeuchi

Abstract

Existing deep models using attributes usually take global features for identity classification and attribute recognition. However, some attributes exist in local position, such as a hat and shoes, therefore global feature alone is insufficient for person representation. In this work, we propose to use the attribute recognition as an auxiliary task for person re-identification. The attributes are recognised from the local regions of mid-level layers. Besides, we extract local features and global features from a high-level layer for identity classification. The mid-level attribute learning improves the discrimination of high-level features, and the local feature is complementary to the global feature. We report competitive results on two large-scale person re-identification benchmarks, Market-1501 and DukeMTMC-reID datasets, which demonstrate the effectiveness of the proposed method.

Keywords: Person re-identification, attribute recognition, local feature.

1. Introduction

Person re-identification (reID) matches the same identity of a query person across non-overlapping cameras. With the development of convolutional neural network, current deep models Lin et al. (2017); Zheng et al. (2017a); Ding et al. (2015); Zhou et al. (2017) can extract strong semantic representation. Most of them use identity labels to train the global feature. However, the global feature may not be discriminative. As shown in Fig.1, from the overall appearance, it is hard to discriminate the persons with clothes of the same color. In these cases, attribute is helpful and can be used as a supplement. For examples in Fig.1, according to the length of lower-body clothes in the first row, the color of shoes in the second row, carrying bags or not in the last row, it is easy to distinguish the target person from the mismatches.

There are two issues about attribute recognition. First, some attribute labels (e.g. wearing glasses and carrying a bag) are extremely unbalanced. Therefore, the model is prone to wrong attribute prediction. Second, the receptive fields of high-level layers are very large, while attributes (e.g. hair, shoes, glasses) usually exist on a small part of a person. In order to alleviate the negative impact on the person descriptor of the error attribute prediction, and retain the contribution of attribute prediction, we address the issues by transforming attribute recognition from the high-level layer to a mid-level layer.

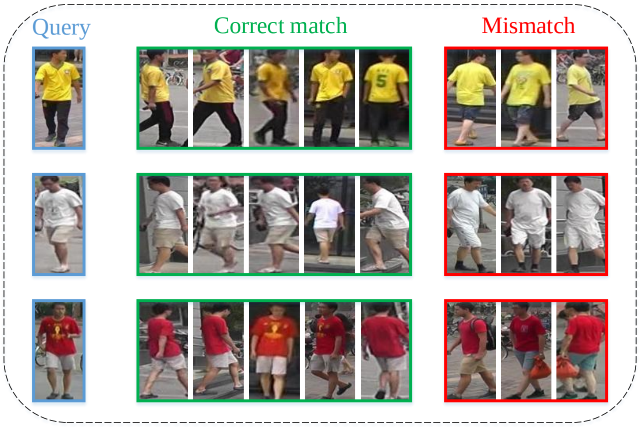


Figure 1: Attributes enhance person reID.

As a result, the attribute learning does not directly work on the global feature for identity recognition. Besides, mid-level feature retains more spatial information than the high-level feature.

Part body feature is also complementary to the global feature. Global feature usually catches the overall appearance of person, while local feature focuses on the salient part body. Therefore, we add a part branch for local feature learning. Under the joint learning of the global feature and the local feature, the feature representation is more discriminative.

The main contributions of this paper includes: 1) We transform attribute recognition into middle layers to promote the discrimination. 2) We create a part branch to capture the complementary local feature. 3) We achieve competitive results on two person reID datasets.

1.1. Related work

Traditional approaches are mainly based on the hand-crafted features. Low-level color and texture are important features to identify the target person. In Li et al. (2013), it divides the image into 24 patches and extracts the color descriptor for each patch. However, the people wearing clothes of similar color are often mistaken for the same identity. In order to mitigate the mismatch, attribute-based descriptors have been proposed for person reID. The ARLTM Liu et al. (2012) uses human-specific attributes as priors to encode targets into semantic topics. In Layne et al. (2014), it proposes major mid-level 'semantic attribute' fusion in synergy with low-level color and texture feature as person description. In Li et al. (2014a), it uses a latent support vector machine to classify clothes attributes to assist person reID.

With the great success of deep learning in computer vision Krizhevsky et al. (2012), Simonyan and Zisserman (2014); Szegedy et al. (2015); He et al. (2016), CNNs have been used for person re-identification Ahmed et al. (2015); Chen et al. (2016); Ding et al. (2015); Li et al. (2014b); Wang et al. (2016). Local information is used as supplement to enhance the discrimination of the global features. In Cheng et al. (2016), it uses the improved triple loss function to learn the fusional feature that connects global feature and 4 part-body features. In Varior et al. (2016), the horizontal divided part images are sent to the long short-term memory networks (LSTMs), and then it employs the contrastive loss to train a siamese network. The spindle network Zhao et al. (2017) first locates 14 human keypoints, then extracts 7 human regions of interest. With human keypoints location, the part-body regions are better aligned between images. Attributes have been investigated for deep models of person re-identification Li et al. (2014a); Su et al. (2016); McLaughlin et al. (2017); Lin et al. (2017). In Su et al. (2016), it proposes the cross-dataset weakly supervised learning method on the principle that the same person should have same attributes. In McLaughlin et al. (2017), a siamese network is proposed for joint attribute recognition and identity verification, which treats the verification loss as the heart of multi-task.

Our work is mostly related to Lin et al. (2017) and Li et al. (2017). In Lin et al. (2017), the global feature from the last convolutional layer is used simultaneously for attribute recognition and identity classification. In our work, we add a local feature from the high layer and attribute recognition is done on a mid-level layer. As demonstrated in the experiments, mid-level attribute learning and the fusion of the local and the global feature will improve the discrimination of the person representation. The JLML model in Li et al. (2017) divides the low-level feature map into 3 horizontal parts. Then, it jointly trains the global branch and part branches. However, this model needs extra parameters and costs more time for part branches. We extract local feature through horizontal pooling for horizontal parts, therefore no extra parameters are needed for the part branch.

2. Our method

In this section, we firstly describe our network architecture in details. Then, we elaborate the loss function used to supervise network.

2.1. network architecture

In the training set of n identities, the training image x_i with the label $\{y_i^{'}, y_{i1}, \ldots, y_{im}\}$, where $y_i^{'} \in [1, \cdots, n]$ represents the identity label, and $\{y_{ij}\}_{j=1}^{m}$ denotes the m attribute labels.

There are two vital components in our network: 1) mid-level attribute recognition; 2) part-body identity learning. We build the joint training model that aims to promote the discrimination of person descriptor. The overall network architecture is depicted in Fig.2. In the dotted box, we apply res50 He et al. (2016) as our backbone network. We duplicate conv5 to add a part branch and use horizontal average pooling to catch the local feature. We utilize 1×1 convolutional layers to decrease channels, because directly pooling results in a very high dimensional local feature (2048×7). Global average pooling is used on the backbone network to learn the global feature. After the two branches , there are two fully-connected (FC) layers of n neurons, where n is the number of identities in training data. Besides, we insert a dropout layer before the FC layer to enhance the sparsity of features. Attribute recognition is connected after res3d. Different from identity classification, most attributes focus on a local unfixed position. We divide the feature map into several horizontal parts so that some attributes are more precisely located as shown in

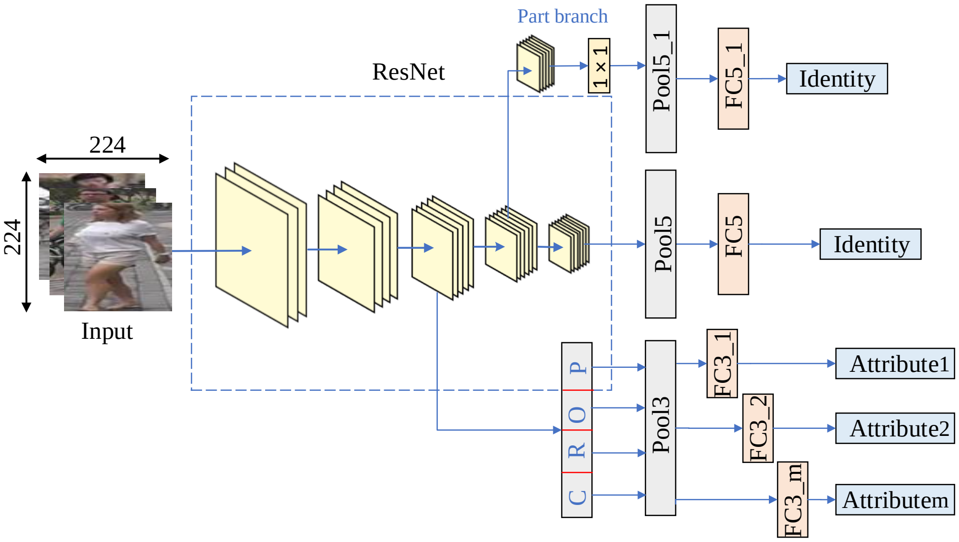


Figure 2: The architecture of our network.

Table 1. We adopt maximum pooling for the attribute learning. After the pooling layers, there are m FC layers for attribute prediction.

2.2. Loss function

The softmax loss function is widely adopted for classification task in convolutional neural network. In the proposed network, for both identity classification and attribute recognition, we use softmax loss to train our model. In each classification task, the predicted probability $\sigma(z) = (\sigma_1(z), \ldots, \sigma_n(z)$ is formulated as below:

$$\sigma_i(z) = \frac{\exp(z_i)}{\sum_{j=1}^{n} \exp(z_j)} \tag{1}$$

In Eq.1, $z_i = W_i^T x + b_i$ is the linear predicted value of the i-th class, where x is the feature vector in pooling layer. For the identity classification task, n is the number of identity in the training set, and for the attribute classification task, n is the type of each attribute. The objective function aims to maximize the value of σ_{y_i}, which is based on the principle of maximum likelihood. Suppose the batchsize is k, the cross entropy loss function is defined as:

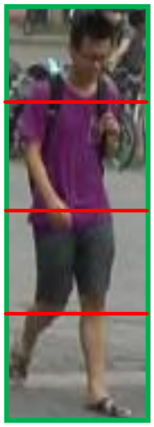
$$L = -\frac{1}{k} \sum_{i=1}^{k} \log(\sigma_{y_i}(z)) \tag{2}$$

Our model is simultaneously supervised by $2 + m$ softmax loss. Therefore, the final loss function is formulated as below:

$$L_{total} = \mu L_{global} + \nu L_{local} + \sum_{i=1}^{m} L_{att\_i} \tag{3}$$

where L_{global} and L_{local} represent the identity classification tasks in the global branch and the local branch, respectively. L_{att_i} denotes one attribute recognition task in conv3 and there are a total of m attribute recognition tasks. In Eq.3, we set two scalars μ and ν to balance the weight of different tasks, and we test the different rates to determine the final value on the validation set.

Table 1: Attributes distribution. Some attributes are distributed over more than one part.



Part 1	age, gender, hair length, sleeve length, carrying bag,carrying backpack, colors of upper-body clothing, wearing hat, carrying handbag
Part 2	age, gender, sleeve length, carrying bag, carrying backpack, carrying handbag, colors of upper-body clothing
Part 3	age, gender, length of lower-body, type of lower-body clothing, colors of lower-body clothing
Part 4	age, gender, length of lower-body, type of lower-body clothing, colors of lower-body clothing, shoe type, color of shoes

3. Experiment

Because we use attribute recognition as an auxiliary task in our model, we mainly perform our experiments on the two person reID datasets, Market-1501 Zheng et al. (2016) and DukeMTMC-reID Zheng et al. (2017a), which contain a large number of samples and detailed attribute labels.

3.1. Datasets

Market-1501 is one of the largest person reID datasets. It consists of 32,668 annotated bounding boxes with 1,501 identities, captured from 6 cameras. Each identity appears in more than one cameras, which ensures the cross-camera match can be conducted. The images are detected by Deformable Part Model (DPM). As the dataset setting, all the annotated bounding boxes are divided into 12,936 training data with 751 identities and 19,732 test data with 750 identities. In testing, for each identity, it randomly picks an image in each camera, so that a total of 3,368 images are used as probe. The market-1501 dataset is collected in the summer. For each identity, there are 27 ID-level attribute labels: male or female, long or short hair, long or short sleeve, long or short lower-body clothing, pants or dress, wearing hat or not, carrying bag or not, carrying backpack or not, carrying handbag or not, 8 colors of upper-body clothing, 9 colors of lower-body clothing and 4 ages (child, teenager, adult or old).

DukeMTMC-reID is a subset of the DukeMTMC Ristani et al. (2016), recently reported by Zheng et al. (2017a). It contains 36,411 bounding boxes, hand-cropped from videos every 120 frames. There are a total of 1,812 identities with 23 attribute labels. These images are divided into 702 identities for training, 702 identities for testing, and 408 identities as "distractor". In the training set, the number of images for each identity is around 10 to 30. But several identities still contain hundreds of images. In the test set, it randomly selects an image for each identity in each camera as Market-1501 does. As a result, the dataset has 16,522 images for training, 2,228 query images and 17,661 gallery images for testing. DukeMTMC-reID dataset is collected in the winter, so the attribute labels is slightly different from the Market-1501 dataset: male or female, boots or other shoes, wearing hat or not, carrying bag or not, carrying backpack or not, carrying handbag or not, dark or light shoes, long or short upper-body clothing, 8 colors of upper-body clothing and 7 colors of lower-body clothing.

3.2. Train

We build our experiment on the commonly used framework Caffe Jia et al. (2014). Our model is fine-tuned on the res50 that was pre-trained from the ImageNet. We use stochastic gradient descent (SGD) algorithm to update the parameters. The initial learning rate is set to 0.01, then reduced by 10 times every 20K iterations. The model is trained up to 50K iterations until convergence. The batchsize is set to a small value of 16. We use the mirror of training data for data augmentation.

The two parameters μ and ν in Eq.3 are important in our experiment. We randomly select 100 images from the training data as validation to determine the appropriate values. Attribute recognition plays a auxiliary role in the training, so we set the weight of each attribute to 0.1. The parameters μ and ν that controls the weight of the identity loss should dominate the loss function. Referring to the APR Lin et al. (2017), when the weight of identity is 8 times that of attribute, the model obtains the best performance in the test phase. Therefore, we control the sum of μ and ν around 8 to verify the performance on the validation set. On Market-1501 dataset, we get the final optimal values with μ = 5 and ν = 4. When the sum of μ and ν is more than 12, attribute recognition does not improve the performance and even degrades it. For DukeMTMC-reID dataset, we use the same values as those on Market-1501.

3.3. Test

In the test phase, in order to verify the effect of different layer features on person reID, we extract the pool5 in the global branch, the pool5_1 in the part branch. It is unsuitable to directly extract the predicted attribute label as the person descriptor. On the one hand, attribute recognition is not accurate due to the unbalanced labels, on the other hand, it is possible that the person with different identities possess the same attribute. Therefore, we only fuse the global feature and local feature as our final person descriptor. However, direct concatenating features of different dimension is unreasonable. The fusional features is prone to high-dimensional feature in similarity metric. To address this problem, we take L2-normalization to the features, so that the difference resulting from large dimension is mitigated. After L2-normalization, the Euclidean distance is equivalent to cosine distance.

Table 2: Results on Market-1501.

Method	rank-1	rank-5	rank-10	mAP
Verif + Identif Zheng et al. (2017b)	79.33	90.76	94.41	55.95
TriNet Hermans et al. (2017)	84.92	94.21	-	69.14
SVDNet Sun et al. (2017)	82.3	-	-	62.1
Spindle Zhao et al. (2017)	76.90	91.5	94.6	-
PDC Su et al. (2017)	84.14	92.73	94.92	63.41
JLML Li et al. (2017)	85.1	-	-	65.50
APR Lin et al. (2017)	84.29	93.20	95.19	64.67
Global feature	86.58	94.48	96.73	68.08
Local feature	86.34	94.00	95.99	65.89
Fusional feature	**87.89**	**94.89**	**97.03**	**70.04**

For similarity metric, we use Euclidean metric to calculate the similarity between a query image and a gallery image. We adopt the Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP) as our evaluation protocols. The CMC reflects the precision, while mAP shows the recall. We use the public evaluation code that available in Zheng et al. (2016).

3.4. Comparisons to Other Methods

We compare our results against the state-of-the-art methods on two datasets, Market-1501 and DukeMTMC-reID. There are two settings on test phase: single-query and multi-query. In the test set, each identity usually has a sequence of ground-truth images. The single-query uses only single query image, while the multi-query using all image of the same identity usually obtains better result. In this work, we only demonstrate the single-query result.

Result on Market-1501. We compare our network with 7 existing models on Market-1501 dataset as shown in Table 2. Most methods use the similar backbone network of ResNet He et al. (2016) except Spindle Zhao et al. (2017) and PDC Su et al. (2017) that design their own networks with inception blocks Szegedy et al. (2015). The first two methods apply metric learning and extract global feature. The SVD Sun et al. (2017) reduces the correlation of global feature with singular vector decomposition. All the next three models utilize local information. They divide the whole people into different parts and extract the fusional feature. Our method was inspired by the APR Lin et al. (2017), which takes the global feature for attribute learning. We transform attribute learning to a mid-level layer and merge the global feature and the part-based local feature. Our model outperforms other models and achieves the advanced results of rank-1 87.89%, rank-5 94.89%, rank-10 97.03% and mAP 70.04%. The performance of global branch and local branch is also competitive due to the contribution of mid-level attribute recognition. Compared with the individual global feature and/or local feature, the fusional feature slightly improves mAP by 1.96% and rank-1 accuracy by 1.31%, which further proves the complementarity of global features and local features.

Table 3: Results on DukeMTMC.

Method	Rank-1	mAP
Basel. + LSRO Zheng et al. (2017a)	67.68	47.13
Basel. + OIM Xiao et al. (2017)	68.1	-
Verif + Identif Zheng et al. (2017b)	68.9	49.3
APR Lin et al. (2017)	70.69	51.88
PAN Zheng et al. (2017c)	71.59	51.51
ACRN Schumann and Stiefelhagen (2017)	72.58	51.96
SVDNet Sun et al. (2017)	76.70	56.80
Global feature	77.73	59.08
Local feature	77.19	57.69
Fusional featrue	**79.67**	**61.48**

Table 4: Results of the global feature with attribute learning from different layers.

feature + attribute recognition	rank-1	rank-5	rank-10	rank-20	mAP
Baseline 1	81.24	91.58	94.02	96.34	60.68
Global + res5c	84.59	93.67	95.53	97.14	64.89
Global + res4f	85.70	94.04	96.26	97.43	65.87
Global + res3d	86.16	94.57	96.50	97.83	66.70

Result on DukeMTMC-reID. As shown in Table 3, we report the rank-1 accuracy and mAP on the DukeMTMC-reID dataset. Our model obtains 79.67% in rank-1 accuracy and 61.48% in mAP, which is better to the other methods. Compared with the APR Lin et al. (2017) that uses attribute recognition and identity classification on the same layer, the results of global feature achieves an increase of 7.04% in rank-1 accuracy and 7.20% in mAP. The fusional feature also outperforms the single feature by 1.94% in rank-1 accuracy and 2.40% in mAP. The results demonstrate that the mid-level attribute learning and the local feature of part-based identity learning indeed improve the performance.

3.5. Ablation study

There are two vital components in our method. On the one hand, we transform attribute recognition from the final convolutional layer to a mid-level layer. On the other hand, we add a part branch to jointly train the local feature and global feature. In order to shed light on the effectiveness of independent element, we design the ablation experiments on Market-1501 dataset.

Effectiveness of mid-level attribute recognition. To prove the contribution of attribute recognition in different layers, we conduct the experiments to test the attribute learning with the global feature and the local feature respectively. Although there are many mid-level layers, attribute recognition is tested in the end of res3d, res4f and res5c, respectively, due to the highest semantic under the same dimensional feature map. In the baseline 1, we test the global feature without attribute. Then, we verify the effect of

Table 5: Results of the local feature with attribute learning from different layers.

feature + attribute recognition	rank-1	rank-5	rank-10	rank-20	mAP
Baseline 2	79.65	91.83	94.69	95.41	56.42
Local + res5c	83.88	93.23	95.27	96.15	61.51
Local + res4f	84.65	93.29	95.90	97.18	63.48
Local + res3d	85.24	93.78	96.05	97.00	65.15

Table 6: The results of fusional feature on Market-1501.

feature	rank-1	rank-5	rank-10	rank-20	mAP
Global	82.10	92.61	94.74	96.70	59.74
Local	79.75	90.94	94.15	96.05	57.32
Fusion	83.82	92.96	95.72	97.06	64.07

attribute learning from the different layers on the global feature, as shown in Table 4. And in the baseline 2, we test the local feature without attribute. Then we individually test the local feature with attribute learning from the different layers, as shown in Table 5. The parameters setting is the same as that in section 4.2.

Compared with the baseline 1 in Table 4, attribute recognition improves the performance of the global feature on Market-1501 dataset. Note that we reproduce the APR network Lin et al. (2017) that directly uses attribute recognition on the global feature. And the result is depicted in the second row (global + res5c). We transform it into lower layers so that some attributes have more precise location. We observe 1.57% performance improvement in rank-1 and 1.81% performance improvement in mAP under this transformation. Therefore, attribute recognition in the mid-level layers indeed promotes the discrimination of global features.

Besides, we demonstrate the contribution of mid-level attribute recognition to local feature. Table 5 evidences the enhancement of local feature due to the transformation of attribute recognition from res5c to res3d. From Table 4 and Table 5, attribute recognition on the conv3 is slightly better than on the conv4 because the feature in conv3 retains more spatial information. We do not test the lower layer (conv2) due to weak semantics.

Effectiveness of fusional feature. In our proposed model, we create a part branch to capture the local identity feature. The feature represented by the part branch is complementary to the global feature. In Table 2 and Table 3, the fusional feature outperforms individual global feature or local feature on the two datasets. To further prove the importance of the part branch, we remove attribute recognition and set the loss weight μ of the local feature the same as the weight ν of the global feature in Eq.3. As depicted in Table 6, the fusional feature obtains 1.72% performance enhancement in rank-1 and 4.33% performance enhancement in mAP, which demonstrates the complementary between the global feature and local feature. It is worth noting that different from most methods that divide the feature map from the conv2, we only used a small branch to obtain local features, which saves the memory and the computation cost.

Furthermore, different attributes perhaps suit differen layers. we try to transform some attributes to different layer. For example, both "age" and "gender" are based on the whole body, so we put "age", "gender" on the higher layer (res4f and res5c). However, this leads to slightly degraded performance. We think this is because there is some correlation between attributes. For example the length of hair is related to the gender. Therefore, under the combination of all attributes on the same layer, attribute recognition improves the discrimination.

4. Conclusion

In this paper, we proposed a deep neural network for person reID. The mid-level attribute recognition enhances the discrimination of the person descriptor. Besides, we combine the complementary global feature and local feature and demonstrate that the fusion indeed improve the model performance. There are still more work to be studied on attribute and fusional feature. For example, how to address the imbalance attribute labels, how to choose the best layer for each attribute and how to distribute the weight of different tasks. We will address these issues in the future work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Project 61175116.

References

Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015.

Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367, 2016.

De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48 (10):2993–3003, 2015.

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.

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

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the Acm*, 60(2), 2012.

Ryan Layne, Timothy M Hospedales, and Shaogang Gong. Attributes-based re-identification. In *Person Re-Identification*, pages 93–117. Springer, 2014.

Annan Li, Luoqi Liu, and Shuicheng Yan. Person re-identification by attribute-assisted clothes appearance. In *Person Re-Identification*, pages 119–138. Springer, 2014a.

Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014b.

Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. *arXiv preprint arXiv:1705.04724*, 2017.

Zhen Li, Shiyu Chang, Feng Liang, Thomas S Huang, Liangliang Cao, and John R Smith. Learning locally-adaptive decision functions for person verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3610–3617, 2013.

Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017.

Xiao Liu, Mingli Song, Qi Zhao, Dacheng Tao, Chun Chen, and Jiajun Bu. Attribute-restricted latent topic model for person re-identification. *Pattern Recognition*, 45(12): 4204–4213, 2012.

Niall McLaughlin, Jesus Martinez del Rincon, and Paul C. Miller. Person reidentification using deep convnets with multitask learning. *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 27, NO. 3, MARCH 2017*, 27(3): 525–539, 2017.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016.

Arne Schumann and Rainer Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1435–1443. IEEE, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, pages 475–491, 2016.

Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3980–3989. IEEE, 2017.

Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *IEEE International Conference on Computer Vision*, pages 3820–3828, 2017.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153. Springer, 2016.

Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1288–1296, 2016.

Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3376–3385. IEEE, 2017.

Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, pages 1116–1124, 2016.

Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 2017a.

Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13, 2017b.

Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. *arXiv preprint arXiv:1707.00408*, 2017c.

Sanping Zhou, Jinjun Wang, Jiayun Wang, Yihong Gong, and Nanning Zheng. Point to set similarity based deep feature learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6, 2017.