Unified Framework for Joint Attribute Classification and Person Re-Identification

Chenxin Sun[1], Na Jiang[1], Lei Zhang[1], Yuehua Wang[2], Wei Wu[1], and Zhong Zhou[1(✉)]

1 State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing, China
2 Department of Computer Science, Texas A&M University – Commerce, Texas, USA
zz@buaa.edu.cn

Abstract. Person re-identification (re-id) is an essential task in video surveillance. Existing approaches mainly concentrate on extracting useful appearance features from deep convolutional neural networks. However, they don't utilize or only partially utilize semantic information such as attributes or person orientation. In this paper, we propose a novel deep neural network framework that greatly improves the accuracy of person re-id and also that of attribute classification. The proposed framework includes two branches, the identity one and the attribute one. The identity branch employs the refined triplet loss and exploits local cues from different regions of the pedestrian body. The attribute branch has an effective attribute predictor containing hierarchical attribute loss functions. After training the identification and attribute classifications, pedestrian representations are derived which contains hierarchical attribute information. The experimental results on DukeMTMC-reID and Matket-1501 datasets validate the effectiveness of the proposed framework in both person re-id and attribute classification. For person re-id, the Rank-1 accuracy is improved by 7.99% and 2.76%, and the mAP is improved by 14.72% and 5.45% on DukeMTMC-reID and Market-1501 datasets respectively. Specifically, it yields 90.95% of accuracy of attribute classification on DukeMTMC-reID, which outperforms the state-of-the-art attribute classification methods by 3.42%.

Keywords: Deep learning, Person re-identification, Attribute classification.

1 Introduction

Person re-identification (re-id) aims at retrieving persons from non-overlapping cameras or different time stamps. Recently, person re-id has been drawing increasing attention from both academia and industry in that it has broad applications in surveillance systems for efficiently preventing and tracking crimes. However, the effects caused by factors like viewpoint variations, occlusion and illumination condition differences potentially make the person re-id an extremely challenging task.

As deep learning arises in the recent years, deep convolutional neural networks have been widely used in person re-id and yielded promising performance [1,2]. However, when being applied to real scenarios, these methods tend to be less effective due to the

lack of detailed cues. In [3], person re-id model is proposed to utilize different parts of the image therefore it can extract regional features containing localized information. The feature maps of different regions of a person appear quite different, which makes the body region alignment of great importance for person re-id. In our re-id framework, we use accurate keypoint locations of a person through keypoint detection to extract desired body regions.

Another common used solution is to exploit person attributes with consideration that the attribute information may contain some domain cues which are identified as the powerful complementary information in the person re-id task [4,5,6,7]. Theoretically, attributes often represent a high level feature of a pedestrian which could be easily missed by approaches based on appearance features. As shown in Fig. 1, people with similar appearance can be easily distinguished by attribute information, which motivated us to study this problem. To solve it, we integrate attribute information into the CNN model for re-id task using our framework.

Fig. 1. Examples of pedestrians in similar appearance with different attribute labels. The attribute labels (e.g., bag vs. handbag, long sleeves vs. short sleeves, etc.) are denoted as discriminative information to distinguish the pedestrians.

The main contributions of this paper are as: (1) A deep neural network incorporating body parts and pose information is proposed. (2) A hierarchical loss guided structure is used to extract meaningful attribute features and consequently to combine the attribute representation with the appearance representation for better re-id. (3) Experiment results on DukeMTMC-reID and Market-1501 datasets demonstrate the effectiveness of the proposed framework. We outperform the state-of-the-art re-id methods in terms of mAP and Rank-1.

2 Related Work

Person re-identification is first introduced and studied by Zajdel et al. [8] in 2005. It is assumed that every individual is associated with unique hidden labels. They design a dynamic Bayesian network to encode the statistical relationships between the features

and the labels of the same identity. Typical traditional person re-identification methods use color or hand-crafted features as feature descriptors. Liao et al. [9] design the Local Maximal Occurrence Representation together with a XQDA metric learning approach for person re-id.

Convolutional Neural Networks have first been used for person re-id by [2,10]. [2] splits the input person images into three horizontal strips processed by several convolutional layers independently. Meanwhile, there are approaches [10,11] which solve re-id problem from the aspect of directly minimizing the feature distance between image pairs or triplets. The Siamese model proposed by Li et al. [10] takes two images as input, directly ending with a same person / different person classification through a deep neural network. Cheng et al. [11] extend this idea and design a similar framework, which processes three images at a time and introduces the triplet loss for metric learning. There are also methods which extract more efficient person features from a tree-structured competitive neural network [3] or different levels of neural network representations [1].

Visual semantic attributes have been investigated in the studies [4,5,6,7,12]. Zhang et al. [4] compute the appearance distances and the attribute distances from two separate models and fuse these two distances together to get the final ranking list. To train unified neural networks, a few methods [5,6,7] use identification and attribute classification loss at the same time to encourage the neural networks to capture both identification and attribute information. However, the information extracted from different domains are difficult to integrate using loss aiming to solve distinct problems. Su et al. [12] propose a weekly supervised multi-type attribute learning algorithm which only uses a limited number of labeled attribute data. In their work, Su et al. employ a three-stage fine-tune strategy to train the model either on attribute datasets or other datasets only labeled with person IDs. The work closest to this paper is [6], in which a combination of re-id and attribute classification losses is used to learn overall representations for person re-id.

3 Proposed approach

We propose a novel deep neural network framework that jointly learns person re-identification and attribute classification, as shown in Fig. 2. Our approach includes an identity branch based on DenseNet-121 and an attribute branch based on ResNet-50 to learn identity and attribute classification respectively. In Fig. 2, the upper part of the framework is the identity branch while the lower part is the attribute branch. At inference time, given as input a person image, we combine identity feature vectors and attribute feature vectors extracted from identity and attribute branch respectively to get the final re-id feature vectors. We then rank the gallery images according to their feature distances to the final representations of the retrieving images. In the following part, we first describe the detail of identity learning framework in Section 3.1 and then the attribute classification structure in Section 3.2.

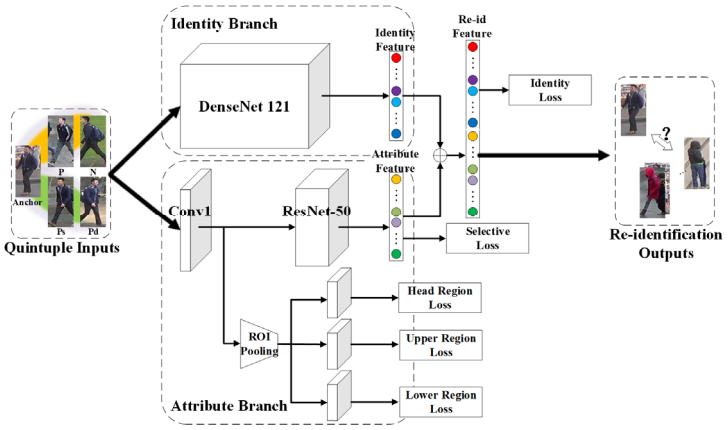

Fig. 2. Overview of our approach. Inputs are quintuples described in Section 3.1.

3.1 Identity Learning Framework

To mitigate occlusions and reduce misalignments, several person re-id studies combine global features with local features which are extracted from certain body parts. Compared with fixed mandatory horizontal strips, accurate body part segmentation can yield more representative local features and greatly eliminate the influence of background. Inspired by such observation, we use the PAFs model [13] to localize fourteen accurate body keypoints and pool three ROI (Region-of-Interest) areas, head, UpperBody and LowerBody, from the feature maps according to the locations of the keypoints. In each forward process, four feature vectors, extracted from the main full image branch and three body part branches, are concatenated to one identity vector which is used for model training, represented by colored rectangle in Fig. 3. Three images on the yellow shadow produce the Triplet loss while three images on the green shadow produce the Orientation loss. Then these two losses are added together to get the identity loss.

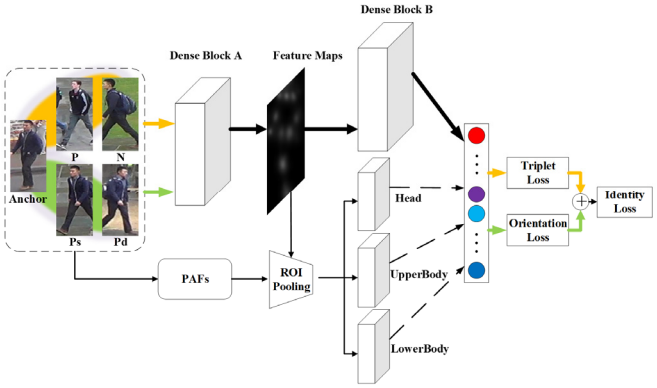

Fig. 3. Identity learning network. Inputs of the convolutional neural network are quintuples including the original image, the positive example, the negative example and two positive examples with same / different orientation, represented by Anchors, P, N, Ps, Pd respectively.

In the training process, we introduce a new orientation-based triplet loss based on the traditional triplet loss [14] in the proposed identity learning model. Concretely, The traditional triplet loss is trained on triplets $\{x_i^a, x_i^p, x_i^n\}$, where x_i^a and x_i^p denote two different images of the same person i, while x_i^n is the third image of a different person. The purpose of triplet loss is to train the network to pull x_i^a closer to x_i^p and push away x_i^n, as formulated as following:

$$Loss_{triplet} = \max\left(d(f(x_i^a), f(x_i^p)) - d(f(x_i^a), f(x_i^n)) + \alpha, 0\right) \quad (1)$$

where $f(\mathrm{x})$ is the feature of the image x, and $d(x, y)$ represents the distance between x and y. α represents the margin between positive pairs and negative pairs.

In our identity learning framework, we argue that we further improve the performance of triplet loss with the pose information. Smaller feature distances between positive samples with the same orientation can be achieved according to the following loss:

$$Loss_{orientation} = \max\left(d(f(x_i^a), f(x_i^{ps})) - d(f(x_i^a), f(x_i^{pd})) + \beta, 0\right) \quad (2)$$

where x_i^{ps} represents the positive sample having the same orientation with anchor sample x_i^a, while x_i^{pd} represents the positive sample having the different orientation. β represents the margin between the same orientation pairs and different orientation pairs. Other symbols in Eq. (2) are the same as the symbols in Eq. (1).

As for the accurate orientation of the images, we use the orientation classification results from the attribute classifier.

The overall loss function for identity learning is formulated as:

$$Loss_{identity} = Loss_{triplet} + \omega * Loss_{orientation} \quad (3)$$

where ω is a weight balancing the two losses of different purposes.

3.2 Attribute Classification

Attributes classifiers are designed to effectively predict the attribute labels and provide meaningful feature vectors to the identity branch for offering complementary information. We dynamically tune training strategies for differentiated phases.

Phase 1. Person attribute classification is formulated as a multitask problem, which requires optimizing all attribute predictors. Suppose we have N training images I_i, (i = 1, …, N) labeled with M attributes $Label_{ij}$, (j = 1, …, M). We need to learn M predictors $\varphi_j(I_i)$ to minimize the difference between the expected output of predictors and the labels, and it can be formulated as follows:

$$\sum_{i=1}^{N}\sum_{j=1}^{M} Loss(\varphi_j(I_i) - Label_{ij}) \quad (4)$$

where Loss (·) in Eq. (4) is the loss function that calculates the difference between the output of each predictor and label; in our experiment, we choose the square loss as loss function.

In the process of training, we observe some attributes have different convergence rates and training difficulties and some attributes like "backpack" and "upwhite" appear more frequently than others. To capture such facts, we follow the approach [14] weighting the attributes in the loss function:

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\lambda_j * Loss(\varphi_j(I_i) - Label_{ij}) \tag{5}$$

where λ_j is the scalar value to weight the importance of attribute j to overall loss function.

Instead of manually tuning the hyper-parameter λ_j using methods like cross validation, we propose an adaptive method to update λ_j every k iterations during training. In each batch, we separate the training images into two parts: the training part and the auxiliary part, all of which are passed through the neural network. We get two kinds of loss vectors from the output of the neural network. But only the loss vector obtained from the training part is used to update the neural network, while the loss vector obtained from the auxiliary part is stored in a data structure $Loss_{[\cdot]}$ used to update the weight vector λ. We formulate the weight update algorithm in Eq. (6) and (7).

$$\lambda = \left[\ \left[Loss_{[n-k:n]} - Loss_{[n-2k:n-k]}\right]_{norm} \cdot \ \left[Loss_{[n-k:n]}\right]_{norm}\ \right]_{norm} \tag{6}$$

$$[\vec{v}]_{norm} = \frac{\vec{v} - v_{min}}{v_{max} - v_{min}} \tag{7}$$

where λ is a M-dim vector, $\cdot$ stands for dot product, $Loss_{[\cdot]}$ is a data structure storing the auxiliary loss vectors, n is the number of losses stored in $Loss_{[\cdot]}$, $Loss_{[b:a]}$ stands for an average loss whose every element is the mean value of the corresponding elements from Loss_a to Loss_b, $[\cdot]_{norm}$ is the normalization function in Eq. (7), v_{min} and v_{max} refer to the minimum and the maximum values in vector $\vec{v}$ respectively, and k is set to 12 with experiential experience in our experiment.

In Eq. (6), the $\left[Loss_{[n-k:n]} - Loss_{[n-2k:n-k]}\right]_{norm}$ factor encourages weights of certain attributes to be larger ones whose current losses change drastically compared to previous losses, while the $\left[Loss_{[n-k:n]}\right]_{norm}$ factor encourages weights of the other kind of attributes to be larger which have not converged. To this end, we keep training our attribute classification network using the weighted loss until convergence, as shown in Fig. 4(a). When we train the attribute classification network with our identity learning network, we use an adaptive strategy to assist the re-id task discussed in Phase 2.

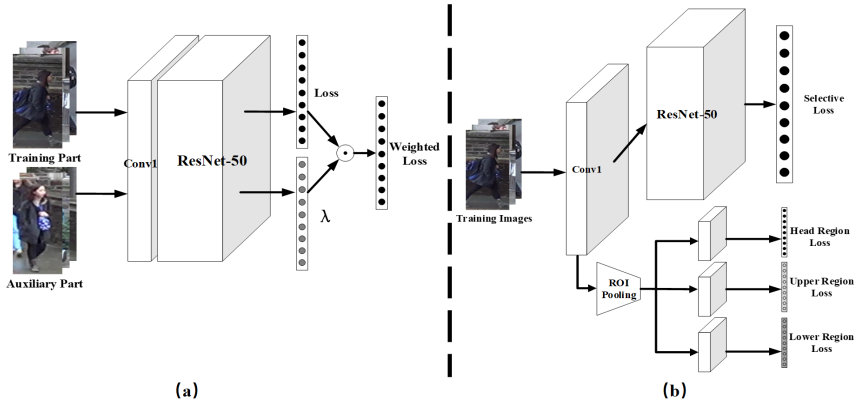

Fig. 4. A figure caption is always placed below the illustration. Short captions are centered, while long ones are justified. The macro button chooses the correct format automatically.

Phase 2. It is noted that attributes in datasets are generally classified into two groups according to whether they can be assigned to certain image regions. The attributes like the color of upper clothes, backpack, and the color of lower clothes, rely on small regions of images rather than the whole images. Based on this observation we design a multi-branch framework for efficient attribute classification which predicts region based attributes respectively, as shown in Fig. 4(b). We initialize the weights of our deep neural network gained by Phase 1 and use the locations of ROI regions in Section 3.1 to pool three regions from the first pooling layer.

Besides, according to the influence of attribute labels on the person re-id task, we choose several attribute labels to train the overall framework using Eq. (4), discarding the prediction layer trained in Phase 1. The selective attribute loss from the main branch together with three losses from region based branches constitute our hierarchical loss.

4 Experiments

4.1 Implementation Details

In our experiments, we choose DenseNet [15] model as our identity branch and ResNet [16] as the attribute classification branch. For the identity branch, it includes a backbone network and three body part subnetworks. They share the weights from the first convolutional layer to the first dense block. We add an ROI pooling layer behind the first dense block to pool three areas from the shared feature maps according to the output of PAFs keypoint estimator. The backbone network and three subnetworks all have four dense blocks with different growth rates. For the attribute branch, the network is designed similarly like the identity branch except that the attribute branch uses proposed hierarchical loss as the objective function.

In the training phase, we firstly use $Loss_{identity}$ in Eq. (3) to train the identity branch and loss in Eq. (5) to train the attribute branch separately until they converge. Secondly, we fix the layers before the pooling1 layer in our attribute branch and copy the layers

after the pooling1 layer to from 3 region based subnetworks. Using proposed hierarchical loss in Phase 2, we train the region based subnetworks and the main attribute branch until convergence. Finally, we concatenate the feature vector extracted from the identity branch and the feature vector obtained from the main part of our attribute branch to get a final re-id feature vector as shown in Fig. 2. Then we calculate the classification loss using this final re-id feature vector, and finetune the whole framework using classification loss and hierarchical loss until convergence.

In the testing phase, we extract a 3048-D feature vectors from the final fused layer. This feature vector has not only identity discriminability but also attribute information. We use this 3048-D feature for person re-id.

4.2 Performance on Attribute Classification

To evaluate the effect of the attribute domain learning, we conduct the attribute classification on DukeMTMC-reID [17] and Market-1501 [18] datasets. In such a way, the identity and attribute labels are obtained for the designed framework.

Table 1. Attribute recognition accuracy on DukeMTMC-reID

Methods	gender	hat	boots	top	back-pack	hand-bag	bag	shoes	up-color	down-color	mean
SVM[24]	77.03	82.24	82.45	87.64	69.59	93.60	83.01	90.05	70.94	68.48	80.50
APR[10]	82.61	**86.94**	**86.15**	88.04	77.28	**93.75**	82.51	**90.19**	72.29	41.48	80.12
Baseline	83.12	81.09	80.52	89.91	76.05	90.06	81.08	81.92	75.54	70.55	80.98
Ours	**88.94**	82.97	80.13	**93.60**	**87.02**	89.60	**91.60**	83.65	**93.94**	**91.84**	**90.95**

Table 2. Attribute recognition accuracy on Market-1501

Meth-ods	gen-der	age	hair	up	down	clothes	back-pack	hand-bag	bag	hat	up-color	down-color	mean
APR[10]	86.45	**87.08**	**83.65**	**93.66**	**93.32**	**91.46**	82.79	**88.98**	**75.07**	**97.13**	73.40	69.91	85.33
Base-line	81.08	85.39	70.49	87.47	84.59	81.51	**86.22**	85.18	67.30	92.10	71.57	71.05	80.33
Ours	**88.94**	84.76	78.26	93.53	92.11	84.79	85.46	88.40	67.28	97.06	**87.50**	**87.21**	**86.98**

In Tables 1 and 2, we compare the attribute recognition accuracy of the proposed method with two state-of-the-art ones, Baseline and APR [5]. Baseline denotes the attribute branch trained by loss in Eq. 4 and Ours represents the attribute classifier fine-tuned by weighted attribute loss in Eq. 5. As shown in the tables, we have achieved competitive results in these two datasets and the proposed framework significantly outperforms the baseline. It is worth noting that the results in [14] are also very competitive with the mean average accuracy of 87.53% and 88.49% on the DukeMTMC-reID and Market-1501 datasets. Our framework achieves 90.95% accuracy on DukeMTMC-reID, outperforming all state-of-the-art methods by 3.42%.

4.3 Performance on Person Re-identification

In this section, we evaluate the performance of our method on the DukeMTMC-reID and Market-1501 datasets.

Table 3. Comparison with the state-of-the-art approaches.

DukeMTMC-reID	Rank-1	mAP	Market-1501	Rank-1	mAP
LOMO+XQDA[9]	30.8	17.0	LOMO+XQDA[9]	43.80	47.78
GAN[17]	67.68	47.13	GAN[17]	79.33	55.95
Loss Embedding[19]	68.90	49.30	Loss Embedding [19]	79.51	59.87
APR[6]	70.69	51.88	ACRN[7]	83.61	62.60
ACRN[7]	72.58	51.96	APR[6]	84.29	64.67
Baseline	67.58	47.46	Baseline	72.50	45.23
Baseline + Triplet	72.33	51.72	Baseline + Triplet	81.32	61.50
Baseline + Improved Triplet	75.72	56.20	Baseline + Improved Triplet	85.88	67.28
Ours	**80.57**	**66.68**	Ours	**87.05**	**70.12**

Table 3 shows the performances of the proposed method comparing to that of several state-of-the-art methods. Baseline represents our identity network without the triplet loss, Baseline + Triplet represents identity network with the original triplet in Eq. 1, Baseline + Improved Triplet represents identity network with proposed triplet loss in Eq. (3) and Ours represents the results of our overall framework in Fig. 2. As shown in Table 3, the Rank-1 accuracy is improved by 7.99% and 2.76%, while the mAP is improved by 14.72% and 5.45% on DukeMTMC-reID and Market-1501 datasets respectively in our overall framework. This result shows the effectiveness of proposed attribute information transferring. With the use of triplet loss and proposed attribute supplementary information, we can observe significant improvement in the final results.

5 Conclusion

In this paper, we have presented a deep convolutional neural framework employing hierarchical attribute information for person re-identification. With the joint learning of the identity and attribute supervision from the same dataset, we invoke information transferring from the attribute domain to the identity domain which is used as supplementary information. According to the evaluation results, the proposed framework shows highly accurate attribute and person re-id comparing to the state-of-the-art methods in the field on two datasets.

Acknowledgment. This work is supported by the Natural Science Foundation of China under Grant No. 61572061, 61472020, 61502020, and the China Postdoctoral Science Foundation under Grant No. 2013M540039.

References

1. Meng, X., Leng, B., Song, G.: A multi-level weighted representation for person re-identification. In: International Conference on Artificial Neural Networks. pp. 80–88. Springer (2017).
2. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: Pattern Recognition (ICPR), 2014 22nd International Conference on. pp. 34–39. IEEE (2014).

3. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: 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. pp. 1077–1085 (2017).
4. Zhang, X., Pala, F., Bhanu, B.: Attributes co-occurrence pattern mining for video-based person re-identification. In: Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on. pp. 1–6. IEEE (2017).
5. Matsukawa, T., Suzuki, E.: Person re-identification using CNN features learned from combination of attributes. In: Pattern Recognition (ICPR), 2016 23rd International Conference on. pp. 2428–2433. IEEE (2016).
6. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y.: Improving person re-identification by attribute and identity learning. arXiv preprint arXiv:1703.07220 (2017).
7. Schumann, A., Stiefelhagen, R.: Person re-identification by deep learning attribute-complementary information. In: Computer Vision and Pattern Recognition Work-shops (CVPRW), 2017 IEEE Conference on. pp. 1435–1443. IEEE (2017).
8. Zajdel, W., Zivkovic, Z., Krose, B.: Keeping track of humans: Have i seen this person before? In: Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on. pp. 2081–2086. IEEE (2005).
9. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2197–2206 (2015).
10. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 152–159 (2014).
11. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: 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. pp. 1335–1344 (2016).
12. Su, C., Zhang, S., Xing, J., Gao, W., Tian, Q.: Multi-type attributes driven multi-camera person re-identification. Pattern Recognition 75, 77–89 (2018).
13. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. vol. 1, p. 7 (2017).
14. He, K., Wang, Z., Fu, Y., Feng, R., Jiang, Y.G., Xue, X.: Adaptively weighted multi-task deep network for person attribute classification. In: Proceedings of the 2017 ACM on Multimedia Conference. pp. 1636–1644. ACM (2017).
15. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. vol. 1, p. 3 (2017).
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016).
17. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned CNN embedding for person re-identification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14(1), 13 (2017)
18. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1116–1124 (2015).
19. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. arXiv preprint arXiv:1701.07717 3 (2017).