Proceedings of Machine Learning Research 95:208-219, 2018 ACML 2018

Discriminative Feature Representation for Person Re-identification by Batch-contrastive Loss

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

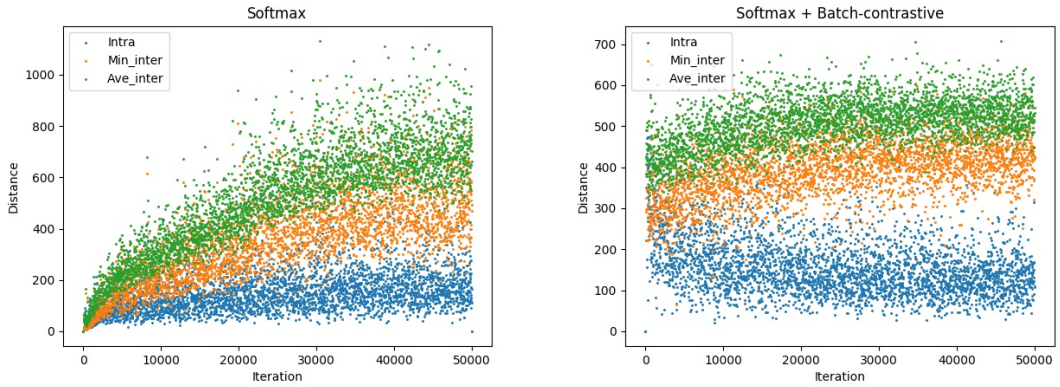
In the past few years, person re-identification (reID) has developed rapidly due to the success of deep convolutional neural networks. The softmax loss function is an important component for learning discriminative features. However, the classifier trained by the softmax loss is difficult to distinguish the hard samples. In this work, we introduce a new auxiliary loss function, called batch-contrastive loss, for person reID to further separate the features of different identities and pulls the features of same identity closer. Furthermore, the proposed loss function does not rely on the pairwise or triplet sampling which is commonly used in the Siamese model. We test our loss function on two large-scale person reID benchmarks, Market-1501 and DukeMTMC datasets. Under the combination of the batch-contrastive loss and the softmax loss, even only employing the generic L2-distance metric, we can achieve competitive results among the state-of-the-arts.

Keywords: Person re-identification, contrastive loss, softmax loss.

1. Introduction

Given an image/video of a person, re-identification (reID) is the process of identifying the same identity taken from non-overlapping cameras. A powerful person re-identification system is useful in some practical applications such as object tracking and video surveillance. However, person reID is affected by reality situation. The system performance degrades due to illumination, pose variations, occlusions and poor quality of the images. In a real video surveillance system, it is hard for cameras to capture human faces clearly. Therefore, person descriptor is usually based on the whole body.

Since the training identities are different from the test identities, we usually first train a person reID model using the training data. Then, we extract the features of the test images using the trained model. Finally, we calculate the similarity between the query images and the gallery images, and obtain the final ranking. We define people with same identity as positive samples and people with different identities as negative samples. Theoretically, the features of positive samples should be close and features of different identities should be separated from each other. Traditional methods mainly use hand-crafted feature such as color and texture. However, models based on these features are not robust when different people have similar appearances or the same person has different poses. In recent years, deep



(a) (b)

Figure 1: Distance between features under the softmax loss and the batch-contrastive loss. The blue (green) dots represent the intra-class (average inter-class) distance, the orange dots denote the minimum inter-class distance.

convolutional neural networks (CNN) have achieved great success in many computer vision fields. The person descriptor learned by the softmax loss function demonstrates strong semantic. Therefore, the classification models Xiao et al. (2016); Wu et al. (2016); Zheng et al. (2016b); Lin et al. (2017); Liu et al. (2016) are widely used to learn discriminative feature for person reID. However, a major drawback of the identification model is the shortage of training data. Besides, identification model does not account for the similarity learning between image pairs. Therefore, it is hard for the single softmax loss function to cluster the hard positive samples and separate hard negative pairs.

In order to further explore the effect of identification model in person reID, we conduct a tiny experiment to observe the intra-class distance and inter-class distance of person features under individual softmax loss. We test the experiment on the Market-1501 Zheng et al. (2016a) dataset. The network structure and parameter settings is the same as the baseline in section 3.3. Because the input images in a batch are randomly selected, calculating the intra-class distance in every iteration is ineffective. In the training phase, we choose the the batch that includes positive pairs. Then we calculate the intra-class distance and all inter-class distance in the batch. This practice roughly reflects distance relation between the features. Under the supervision of individual softmax loss function, we plot the intra-class distance, the minimum inter-class distance and the average inter-class distance as shown in Fig.1(a). From the green dots and blue dots, the intra-class distance is obviously smaller than the average inter-class distance, which demonstrates the effective of softmax loss function. However, from the intersection between orange dots and blue dots, there are still some hard negative pairs closer than the intra-class distance. Considering that the minimum inter-class distance is only chosen in a batch, there are more hard negative pairs when extending to all training samples. Therefore, the individual softmax loss function can roughly separate from different classes, but it is difficult to distinguish hard inter-class examples.

To alleviate this problem, metric learning is introduced to enhance the discrimination of person descriptor. The Siamese network consists of two/three similar networks that are supervised by the contrastive/triple distance loss Yi et al. (2014); McLaughlin et al. (2017); Ding et al. (2015); Cheng et al. (2016); Zheng et al. (2016b). The two losses are proved to contribute to the discrimination by pulling the features of the same identity close and pushing the features of different identities far away from each other. In Zheng et al. (2016b), an identification model and a verification model are combined. The identification model makes the features from different classes separable. The verification model forces the features of the same identity closer and the features of different identities far apart. The triple loss Wang et al. (2016) makes the distance of a positive pair smaller than that of a negative pair by a suitable margin. However, both pairwise and triplet samples grow explosively as the training data increases. It is hard for the Siamese network to achieve optimal because it only uses weak labels (same or different) and does not take all the annotated information into consideration. Besides, it takes more time and requires more memory due to the shared branches in the training phase. Some works Ding et al. (2015); Shi et al. (2016) investigate how to sample the hard triplets and pairs. In Lin et al. (2017); McLaughlin et al. (2017), auxiliary attribute classification tasks are utilized to enhance the discrimination of features. However, manual attribute labelling usually costs a lot and some person reID datasets do not possess attribute labels.

In this paper, we design a new loss function that aims to further reduce the distance of intra-class features while enlarging the distance of the inter-class features. Note that the proposed loss function is based on a batch and does not require to sample the image pairs or triplets as the network input. In a randomly sampled batch, almost all image pairs in the batch are negative samples, so we utilize the negative pairs to increase the inter-class distance. However, there are few positive samples to penalize the intra-class loss due to the random sampling. To address this problem, we introduce the center loss Wen et al. (2016) to balance the inter-class distance loss. The center loss is firstly proposed in face recognition. It learns a center for each class and at the same time forces the feature within the same class close to the center. As a result, the features of the same identity indirectly approach to each other. Individual center loss only focus on the intra-class distance and does not consider the inter-class distance. Therefore, we incorporate the center loss into the contrastive loss. The proposed loss function is based on a batch, so we call it batch-contrastive loss.

The main contributions of this paper are three folds: 1) We propose a new loss function that enforces the features of the same identity into a cluster while separating the inter-class features from each other. Different from the Siamese network, our loss function is based on a batch so that we do not need to sample image pairs or triplets as the network input; 2) The proposed batch-contrastive loss make full use of the image pairs in every batch. Besides, it has a hard pairs mining process; 3) With the joint supervision of the softmax loss and the batch-contrastive loss, we achieve competitive results on two large public Re-ID datasets, Market-1501 Zheng et al. (2016a) and DukeMTMC-reID Zheng et al. (2017a).

2. Proposed method

In this section, we first describe the network architecture and the softmax loss in details and then elaborate the proposed batch-contrastive loss.

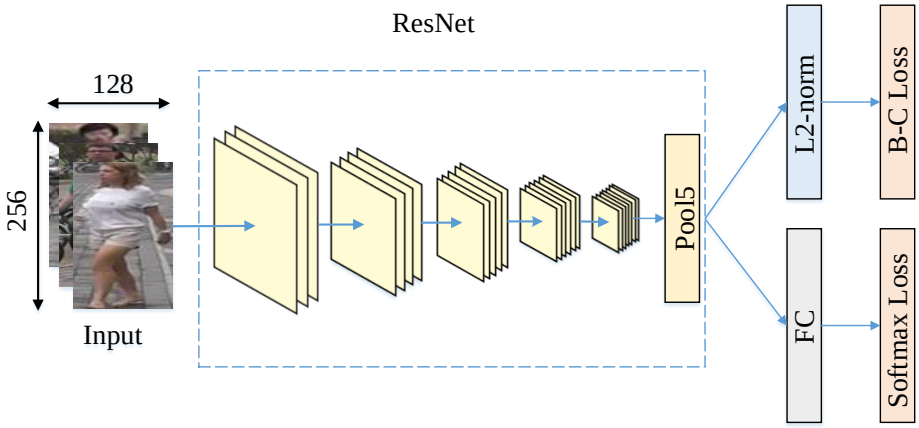


Figure 2: The architecture of our network.

2.1. Network Architecture

The network architecture is depicted in Fig.2. In the dotted box, it consists of the layers from conv1 to pool5 of ResNet50 He et al. (2016). After the pool5, we add a dropout layer to avoid overfitting. The fully-connected(FC) layer has n neurons, where n is the number of identities in the training set. Our network is simultaneously supervised by the softmax and the batch-contrastive (B-C) loss. Furthermore, we insert a normalization layer before the batch-contrastive loss for two benefits. On the one hand, the cosine distance is usually superior to the Euclidean distance in similarity metric, and when the vector is L2-normalized, the batch-contrastive loss calculates the Euclidean distance that is equivalent to the cosine distance. On the other hand, after L2-normalization, the inter-class distance is limited to a small value of 1, which avoids the enormous contrastive loss resulted from random initialization overwhelming the effect of the softmax loss.

In the case of data shortage, the dropout Srivastava et al. (2014) plays an important role in alleviating overfitting, typically working on FC layer. It randomly discards the neural units to enhance sparsity of the network. We utilize the dropout strategy on pool5 layer because the feature map (1×1) of pool5 is similar to the unit of FC layer, which makes the discard work on the channel.

2.2. Loss Function

Softmax loss. The softmax loss function is commonly adopted for classification task in the convolutional neural network Krizhevsky et al. (2012); Simonyan and Zisserman (2014); Szegedy et al. (2014); He et al. (2016). Given a batch training data of m images, the predicted probability $\sigma(z) = (\sigma_1(z), \ldots, \sigma_n(z))$ for each sample is calculated 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 result of the ith class, where x is the dropout vector extracted from the dropout layer, and W_i and b_i are the parameters in FC layer. The

purpose of the softmax loss is to maximize the value of σ_{y_i}, which is based on the principle of maximum likelihood, therefore the cross entropy loss function is formulated as:

$$L_s = -\frac{1}{m}\sum_{i=1}^{m}\log(\sigma_{y_i}(z)). \tag{2}$$

Batch-contrastive loss. In a batch of m images, there are a total of $\frac{1}{2}\times m\times(m-1)$ image pairs. However, most pairs are easily distinguished by the softmax loss function as shown in Figure.1(a). It makes no sense to further increase the distance of these easy negative pairs. Therefore, for each data, we calculate all inter-class distance in the batch, and then choose the minimal distance to create the hard negative pairs. This is a hard negative mining process. The hard negative pairs is better to reflect the difference between samples. In addition, a few positive pairs appear in some batches. We directly penalize the intra-class distance to further pull the feature of positive pairs close. The batch-based contrastive loss function is formulated as below:

$$\mathcal{L}_b = \frac{1}{2}\sum_{i=1}^{m}\left(\sum_{j=1}^{m}\left(\delta(y_i=y_j)\left\|x_i-x_j\right\|_2^2\right)+\max\left(0,\tau-\min_{j,y_i\neq y_j}\left\|x_i-x_j\right\|_2^2\right)\right). \tag{3}$$

In Eq.3, x represents the feature in pool5 and y is the label (ID) where the index i,j denotes the ith and the jth image in the batch, respectively. $\|\cdot\|_2^2$ denotes the Euclidean distance. δ is the indicator function, that is $\delta(\cdot)=1$ if the condition in the brackets is true, and otherwise $\delta(\cdot)=0$. τ represents the margin controlling the minimum hard negative distance. When inter-class distance is larger than τ, the inter-class loss is set to 0. Eq.3 takes all image pairs into consideration. The first part penalizes the intra-class distance loss, and the second part denotes the inter-distance loss. However, there are only few even no positive pairs in every batch due to the random sampling, which results in the intra-class loss getting overwhelmed by the inter-class loss significantly. In order to balance the two mutually exclusive distance losses, we naturally introduce center loss to further reduce the intra-class distance.

The center loss Wen et al. (2016) aims to minimize the distance between the features and their centers. It can be formulated as below:

$$\mathcal{L}_c = \frac{1}{2}\sum_{i=1}^{m}\left\|x_i-c_{y_i}\right\|_2^2. \tag{4}$$

In Eq.4, c_{y_i} represents the center of class y_i. Ideally, in each iteration, we should calculate the new center c_{y_i} using all training samples with the label y_i, as the update of weights results in the distribution of features changing. However, computing all new centers will lead to enormous computation, which is considerably inefficient.

The network updates weights based on the batch data. In this case, we use a small number of samples to approximate the center. The centers c_{y_i} are learned according to the samples within the mini-batch as follows:

$$\Delta c_j = \frac{\sum_{i=1}^{m}\delta(y_i=j)\cdot(c_j-x_i)}{1+\sum_{i=1}^{m}\delta(y_i=j)}, \tag{5}$$

$$c_{j_{\text{new}}} = c_{j_{\text{old}}} + \eta \Delta c_j, \tag{6}$$

where η denotes the learning rate of centers.

Finally, we combine the batch-based contrastive loss with the center loss to enhance the discrimination. The batch-contrastive loss is finally given in Eq.7.

$$\mathcal{L}_{B-C} = \mathcal{L}_b + \mathcal{L}_c. \tag{7}$$

In the process of backward propagation, according to the batch-contrastive loss function, the gradients of $\mathcal{L}_{B-C}$ with respect to x_i is calculate as below:

$$\begin{aligned} \frac{\partial \mathcal{L}_{B-C}}{\partial x_i} &= \frac{\partial \mathcal{L}_b}{\partial x_i} + \frac{\partial \mathcal{L}_c}{\partial x_i} \\ &= \sum_{j=1}^{m} (\delta(y_i = y_j)(x_i - x_j)) - (x_i - x_{min_i}) + x_i - c_{y_i}. \end{aligned} \tag{8}$$

In Eq.8, the hard negative sample x_{min_i} satisfies $\|x_i - x_{min_i}\|_2 < \tau$ and $\|x_i - x_{min_i}\|_2 \leq \|x_i - x_j\|_2$ for all j with $y_i \neq y_j$.

By jointly using the batch-contrastive loss and softmax loss, the final loss function is given as follows:

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_{B-C}. \tag{9}$$

In Eq.9, the parameter λ is to make tradeoff between the softmax loss and batch-contrastive loss. To learn the discriminative pedestrian descriptor, the parameters updating algorithm is formulated as below.

Algorithm 1 The network updating algorithm.

Input:

The randomly selected images in current batch;
The parameters in the backbone network W;
The center in the batch-contrastive loss, c;
The learning rate η, loss weight λ and the initial $t = 0$.

Output:

The updated parameters, W.

while *iteration* $<$ *max_iteration* **do**
  $t \leftarrow t + 1$
  Compute the forward propagation loss, $\mathcal{L}$.
  Compute the backward propagation gradient, $\frac{\partial \mathcal{L}^t}{\partial x_i} = \frac{\partial \mathcal{L}_s^t}{\partial x_i} + \lambda \cdot \left(\frac{\partial \mathcal{L}_b^t}{\partial x_i} + \frac{\partial \mathcal{L}_c^t}{\partial x_i} \right)$
  Compute the new centers $c_{new} = c_{old} + \eta \cdot \Delta c$
  Compute the new parameters $W_{new} = W_{old} + \eta \cdot \frac{\partial \mathcal{L}^t}{\partial x_i} \cdot \frac{\partial x_i}{\partial W_{old}}$
end while

3. Experiments

Since the auxiliary batch-contrastive loss is based on intra-class distance and inter-class distance, we mainly verify the proposed method on two large-scale person ReID datasets Market-1501 and DukeMTMC-reID, which contain multiple samples for each identity.

3.1. Datasets

Market-1501 consists of 32,668 annotated bounding boxes with a total of 1,501 identities, captured by 6 cameras from different view points. The images are detected by Deformable Part Model (DPM). Each identity appears in at least two cameras, which ensures the cross-camera match can be conducted. Following the setting in Zheng et al. (2016a), 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 set, for each identity, it randomly picks an image in each camera, so that a total of 3,368 images are used as probe. In the gallery set, it contains"good" images, "junk" images that have no influence to the ReID accuracy, and "distractor" of false alarms.

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, cropped from videos every 120 frames. There are a total of 1,812 identities that 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.

3.2. Implementation

Training. We perform our experiments on the deep learning framework Caffe Jia et al. (2014). The ResNet50 is pre-trained with the ImageNet Krizhevsky et al. (2012). We fine-tune the ResNet50 model using stochastic gradient descent(SGD). Different from conventional input size of 224×224, it is more suitable to resize the pedestrian image to 256×128. In fact, this transformation subtly improves the performance. We use a small batch size of 16 because the larger batch size is prone to fall in local optimum. The initial learning rate is set to 0.01 and then dropped by 10 times every 20K iterations. The model is trained up to 50K iterations until convergence. The parameter λ is used to balance the two loss function. we randomly choose 100 images from training data to verify the the performance of different values. When $\lambda = 10$, the batch-contrastive loss quickly converges to a small value. However, the performance decreases due to excessive weight. And when λ is less than 0.1, it only results in small improvement. Therefore, we set $\lambda = 1$ in the final loss function. We choose a very high dropout rate of 0.9 for our model.

Testing. In the test stage, for each image in the probe and the gallery, the model extracts 2048-dimensional feature from pool5. After L2-normalization, we use the normalized feature as the pedestrian descriptor. For each query image, we calculate the distance between the query image and all gallery images using Euclidean distance metric. Then we rank the distance to obtain the final result. For the evaluation protocols, we choose the widely adopted Cumulative Matching Characteristic (CMC) curve and the mean average precision(mAP) Zheng et al. (2016a), which measure the precision and recall, respectively. The public evaluation code is available in Zheng et al. (2016a, 2017a).

Table 1: Results on Market-1501. "-" denotes the results are not reported. "SQ" represents single-query.

Method(SQ)	rank-1	rank-5	rank-10	mAP
SpindleNet Zhao et al. (2017)	76.90	91.50	94.60	-
GAN Zheng et al. (2017a)	78.06	-	-	56.23
SVDNet Sun et al. (2017)	82.3	-	-	62.1
PDC Su et al. (2017)	84.14	92.73	94.92	63.41
APR Lin et al. (2017)	84.29	93.20	95.19	64.67
JLML Li et al. (2017)	85.1	-	-	65.5
Verif + Identif Zheng et al. (2016b)	79.51	90.91	94.09	59.87
TriNet Hermans et al. (2017)	84.92	94.21	-	**69.14**
Baseline	83.34	93.79	95.93	63.20
Ours	**86.40**	**94.45**	**96.48**	67.64

3.3. Baseline

In order to verify the contribution of the batch-contrastive loss, we build a baseline with the individual softmax loss. We fine-tune the ResNet50 on the two ReID datasets. The parameters setting is the same as that in section 3.2.

3.4. Performance on Market-1501 and DukeMTMC-reID

We compare the results of our method with several other algorithms on the two datasets. There are two settings for evaluation: single-query and multi-query. The single-query only uses the feature of one query image, while the multi-query utilizes the feature mean of multiple query images of the same identity from the same camera. The multi-query usually achieve a better result because intra-class variation is taken into account. In this paper, we only show the single-query result. We repeat the experiment 5 times, and use the average performance as the final result.

Results on market-1501. As shown in Table 1, we compare our method with the baseline and 8 existing models. We achieve the competitive results of rank-1 = 86.40%, rank-5 = 94.45%, rank-10 = 96.48% and mAP = 67.64%. Compared with the baseline that uses only softmax loss, our method outperforms the baseline by 3.06% in rank-1 accuracy and 4.44% in mAP, which demonstrates the effectiveness of the proposed loss function. In the Siamese network, the verification loss Zheng et al. (2016b) achieve rank-1 79.51% and the triplet loss Hermans et al. (2017) obtains the result of rank-1 84.92%. However, both methods need to carefully sample the image pairs or triplets. It should also be noted that our method also outperforms the Siamese network in rank-1 accuracy. **Result on DukeMTMC-reID.** In Table 2, we report the rank-1 accuracy and mAP on the DukeMTMC-reID dataset because most methods only shows the rank-1 accuracy and mAP. Our model achieves 78.19% in rank-1 accuracy and 58.64% in mAP, which is superior to other current algorithm. Compared with the of result APR Lin et al. (2017) and ACRN Schumann and Stiefelhagen (2017) that use extra attribute information to enhance the discrimination of feature, our

Table 2: Results on DukeMTMC-reID.

Method(SQ)	Rank-1	mAP
Verif + Identif Zheng et al. (2016b)	68.90	49.30
APR Lin et al. (2017)	70.69	51.88
PAN Zheng et al. (2017b)	71.59	51.51
ACRN Schumann and Stiefelhagen (2017)	72.58	51.96
FMN Ding et al. (2017)	74.51	56.88
SVDNet Sun et al. (2017)	76.70	56.80
Baseline	73.31	53.25
Ours	**78.19**	**58.64**

model gain an increase of about 5.61% in rank-1 accuracy and 6.68% in mAP. Compared with the baseline, the auxiliary batch-contrastive loss leads to the improvement of 4.88% in rank-1 accuracy and 5.39% in mAP, which further verifies the efficiency of the proposed method. To shed light on effect of the proposed loss function, we repeat the tiny experiment in the introduction under the joint training. We plot the results in Fig.1(b). Compared with Fig.1(a) that use individual softmax loss, although there are a few blue dots still mixed in orange dots, it is obvious that most orange points is gradually separated from blue points. That demonstrates that the intra-class distance is less than the minimum inter-class distance. This transformation further indicates that our batch-contrastive loss function indeed separates the hard negative pairs far apart.

4. Conclusion

In this paper, we proposed an efficient loss function to enhance the discrimination of pedestrian feature in person re-identification. The auxiliary batch-contrastive loss function computes the center distance loss, while penalizing the distance of the hard negative pairs. An advantage of our method is that we do not need to sample the image pairs. Besides, the proposed loss function has a process of hard samples mining. Finally, we report the performance on two large-scale person re-identification datasets, which proves the effectiveness of the batch-contrastive loss.

Acknowledgments

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

References

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.

Guodong Ding, Salman Khan, Zhenmin Tang, and Fatih Porikli. Let features decide for themselves: Feature mask network for person re-identification. *arXiv*, 2017.

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.

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

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.

Jiawei Liu, Zheng Jun Zha, Q. I. Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. Multi-scale triplet cnn for person re-identification. In *ACM on Multimedia Conference*, pages 192–196, 2016.

Niall McLaughlin, Jesus Martinez del Rincon, and Paul C. Miller. Person reidentification using deep convnets with multitask learning. *IEEE Trans. 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.

Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z Li. Embedding deep metric for person re-identification: A study against large variations. In *European Conference on Computer Vision*, pages 732–748. Springer, 2016.

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

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2014.

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.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

Shangxuan Wu, Ying-Cong Chen, Xiang Li, An-Cong Wu, Jin-Jie You, and Wei-Shi Zheng. An enhanced deep feature representation for person re-identification. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 34–39. IEEE, 2014.

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, 2016a.

Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person re-identification. *arXiv preprint arXiv:1611.05666*, 2016b.

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. Pedestrian alignment network for large-scale person re-identification. *arXiv preprint arXiv:1707.00408*, 2017b.