A DISCRIMINATIVELY LEARNED FEATURE EMBEDDING BASED ON MULTI-LOSS FUSION FOR PERSON SEARCH

Hong Liu, Wei Shi, Weipeng Huang, Qiao Guan

Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, China
{hongliu, pkusw, wepon, guanqiao}@pku.edu.cn

ABSTRACT

Person search is a challenging task that requires to address pedestrian detection and person re-identification simultaneously. Though significant progress has been made in detection and re-identification respectively, the similar appearances of persons, pedestrian misdetections and false alarms still have adverse effects on person search. To this end, an improved end-to-end person search network with multi-loss is proposed to jointly optimize detection and re-identification. Firstly, a pretrained network is designed to obtain proper initial state for the whole training network. Then, to enhance the person search model, an improved online instance matching (IOIM) loss is proposed by hardening the distribution of labeled identities and softening the distribution of unlabeled identities. Finally, considering the intra-class compactness of features learned by center loss, the IOIM loss is combined with center loss by the proposed multi-loss fusion strategy, which can learn more discriminative feature embeddings. Experimental results on two challenging datasets CUHK-SYSU and PRW demonstrate our approach significantly outperforms the state-of-the-arts.

Index Terms— Person Search, Multi-Loss, Feature Embedding

1. INTRODUCTION

Person search aims at localizing and matching query persons from the whole monitoring gallery image without relying on the annotations of pedestrian candidate boxes [1]. It has wide range of applications in areas such as video analysis [2], intelligent surveillance [3], and other systems [4]. Many person search methods [1, 5, 6, 7] are mainly composed of two components: a pedestrian detector [8] to determine the locations of pedestrian candidates, and a person re-identification algorithm [9, 10] to re-identify the detected candidates. Although significant progress has been made in both pedestrian detection and person re-identification, person search still remains a challenging problem due to the similar appearances of persons, pedestrian misdetections and false alarms.

Relation to prior work: In recent years, Convolutional Neural Networks (CNNs) have shown the potential for learning feature embeddings and similarity metrics in person search. Most existing methods [1, 6, 7] utilize an end-to-end CNNs model to jointly optimize both pedestrian detection and person re-identification. The feature embeddings of identities in training set are learned by training the end-to-end CNNs model. However, each identity contains only several samples in person search. If we directly train the end-to-end model with the weights drawn from Gaussian distributions, it is difficult to learn the discriminative feature embeddings for different identities. To this end, a pretrained model, which distinguishes the cropped identities and backgrounds, is proposed to provide proper initial state for the whole end-to-end training model.

A recent trend towards learning features is to reinforce CNNs with more discriminative information [11, 12]. One way is to learn feature embeddings with the classification loss [13, 14]. The softmax loss [7, 15, 16], a classical classification loss, has been investigated intensively due to its simplicity and probabilistic interpretation. Wen *et al*. designed center loss [17] to further minimize the intra-class distance by penalizing the distances between the features of samples and their centers. Specifically, since the number of identities is very large in person search, the model with the softmax loss is difficult to converge. Xiao *et al*. [1] proposed the non-parametric online instance matching (OIM) loss to accelerate the convergence of the person search model. They treated all unlabeled identities as one class, and set a fixed temperature to harden the distributions of both labeled and unlabeled identities. However, the hard distribution makes the unlabeled identities of the same class more separated in feature space, which is adverse to the convergence of the model. In this paper, the improved online instance matching (IOIM) loss, which hardens the distribution of labeled identities and softens the distribution of unlabeled identities, is proposed to enlarge the differences of labeled identities and minimize the differences of unlabeled identities. To further minimize the intra-class distance of identification feature embeddings, we propose a multi-loss fusion strategy by combining the IOIM loss and center loss. In this way, the learned feature embeddings of each class are more centralized.

This work is supported by National Natural Science Foundation of China (NSFC, No.U1613209, 61340046, 61673030), Natural Science Foundation of Guangdong Province (No.2015A030311034), Scientific Research Project of Guangdong Province (No.2015B010919004), Specialized Research Fund for Strategic and Prospective Industrial Development of Shenzhen City (No.ZLZBCXLJZI20160729020003), Scientific Research Project of Shenzhen City (No.JCYJ20170306164738129), Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (No.ZDSYS201703031405467).

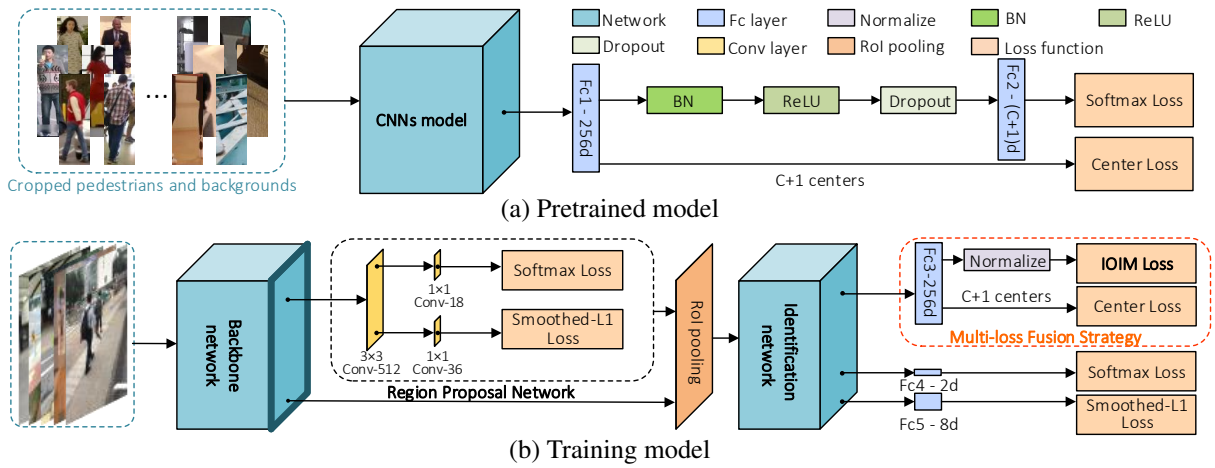

Fig. 1. Overall framework of improved end-to-end person search network (**Best viewed in color**).

2. PROPOSED METHOD

2.1. Method Overview

Let K denote the gallery size which refers to the number of the gallery images. Firstly, the improved person search network is trained with the proposed multi-loss fusion strategy. Then, all gallery images $\boldsymbol{G} = \{\boldsymbol{G}_1, ..., \boldsymbol{G}_k, ..., \boldsymbol{G}_K\}$ are fed into the trained person search model to detect pedestrian candidates $\boldsymbol{P} = \{\boldsymbol{P}_{G_1}, ..., \boldsymbol{P}_{G_k}, ..., \boldsymbol{P}_{G_K}\}$ and extract their feature embeddings. Let $\boldsymbol{f}(\boldsymbol{P}^i_{G_k})$ represent the feature embedding of the i-th pedestrian candidate in the $\boldsymbol{G}_k$-th gallery image. The query person $\boldsymbol{Q}$ is fed into the trained person search model without the region proposal network to extract the feature embedding $\boldsymbol{f}(\boldsymbol{Q})$. Finally, the cosine distances between $\boldsymbol{f}(\boldsymbol{Q})$ and $\boldsymbol{f}(\boldsymbol{P}^i_{G_k})$ are calculated to search the query person $\boldsymbol{Q}$ in the gallery $\boldsymbol{G}$.

2.2. Improved Person Search Network

Fig. 1 depicts that our framework contains two major phases. In pretraining phase, the pretrained model is designed to obtain proper initial state for training phase. In training phase, the training model is used to learn the feature embeddings for query person $\boldsymbol{Q}$ and pedestrian candidates $\boldsymbol{P}$ in the gallery $\boldsymbol{G}$.

(1) Pretraining phase. The pretrained model is illustrated in Fig. 1 (a). We crop the ground truth bounding boxes of all training identities from the training set, and randomly sample the same number of background boxes. The number of training identities is C, and all background boxes are treated as one class. All of them are resized to 224×224, and fed into the CNNs model. The CNNs model is connected with a 256 dimensional fully-connected (Fc) layer which is followed by two branches. One is directly connected with center loss [17] which learns C+1 centers to make the features of intra-class much closer, and the other is followed by a batch normalization (BN) layer, a ReLU layer, a dropout layer, a C+1 dimensional Fc layer, and softmax loss. The model pretrained with these two loss functions is utilized to initialize the training model.

(2) Training phase. As depicted in Fig. 1 (b), the training model contains three major modules: the backbone network, the region proposal network and the identification network. The structures of the three modules are described as follows:

Backbone network. The backbone network adopts the front part of the pretrained CNNs model to learn features for pedestrian detection and identification.

Region proposal network. The region proposal network [18] is utilized to detect the pedestrian candidates in the whole monitoring image. According to [1], a $512 \times 3 \times 3$ convolutional (Conv) layer is connected to the feature maps obtained by the backbone network. An $18 \times 1 \times 1$ Conv layer followed by softmax loss is used to predict the scores of pedestrians candidates or backgrounds, and a $36 \times 1 \times 1$ Conv layer with smoothed-L1 loss is used to predict the locations of pedestrian candidates. According to the predicted locations, all feature maps of pedestrian candidates are cropped and converted with a fixed size by the region of interest (RoI) layer.

Identification network. The identification network composed of the rest of the pretrained CNNs model is employed to learn feature embeddings of pedestrian candidates.

Overall, the end-to-end model is trained with several loss functions. A two dimensional Fc layer with softmax loss is deployed to eliminate backgrounds. An eight dimensional Fc layer with smoothed-L1 loss is utilized to refine the locations of pedestrian candidates. To further distinguish different identities, we propose a multi-loss fusion strategy to learn discriminative identification feature embeddings.

2.3. Multi-loss Fusion Strategy

The online instance matching (OIM) loss [1] is a non-parametric function which utilizes a lookup table $\boldsymbol{V} \in \mathbb{R}^{D \times C}$ to store the features of all labeled identities, and a circular queue $\boldsymbol{U} \in \mathbb{R}^{D \times Z}$ to store the features of those unlabeled identities in recent mini-batches. The OIM loss is defined as:

$$L_{OIM} = -\sum_{i=1}^{M} log\left(\frac{e^{\boldsymbol{v}_t^T \boldsymbol{x}_i/\tau}}{\sum_{j=1}^{C} e^{\boldsymbol{v}_j^T \boldsymbol{x}_i/\tau} + \sum_{k=1}^{Z} e^{\boldsymbol{u}_k^T \boldsymbol{x}_i/\tau}}\right), \quad (1)$$

where M is the number of training data in a batch, and $\boldsymbol{x}_i$ is the i-th normalized 256 dimensional feature with the label t

($t \in [1, C]$, C is the number of training identities). The term $\boldsymbol{v}_t^T \boldsymbol{x}_i$, in which $\boldsymbol{v}_t$ is the t-th column of $\boldsymbol{V}$, denotes the score classified as the t-th labeled identities, and $\boldsymbol{u}_k^T \boldsymbol{x}_i$ represents the score classified as the k-th unlabeled identity. The item $\boldsymbol{u}_k$ is the k-th column of $\boldsymbol{U}$ ($k \in [1, Z]$, and Z is the queue size). The temperature parameter τ is related to the probability distribution over different classes [19].

When τ is relatively smaller, the differences of $\boldsymbol{u}_k^T \boldsymbol{x}_i$ with different k are larger, which results in the harder probability distribution over different unlabeled identities. Therefore, different unlabeled identities are more discriminative. The terms $\boldsymbol{v}_t^T \boldsymbol{x}_i$ and $\boldsymbol{v}_j^T \boldsymbol{x}_i$ have similar principle to $\boldsymbol{u}_k^T \boldsymbol{x}_i$. In OIM loss function, the same τ is set for $\boldsymbol{v}_t^T \boldsymbol{x}_i$, $\boldsymbol{v}_j^T \boldsymbol{x}_i$ and $\boldsymbol{u}_k^T \boldsymbol{x}_i$, which aims to make both labeled and unlabeled identities more discriminative. However, all unlabeled identities are treated as one class in OIM loss function. The differences among unlabeled identities in feature space will increase the intra-class errors. Therefore, we modify OIM loss by using a smaller τ_1 ($\tau_1 < 1$) to harden the distribution over labeled identities and a larger τ_2 ($\tau_2 > 1$) to soften the distribution over unlabeled identities, generating the improved online instance matching (IOIM) loss. The mathematical expression of IOIM loss is:

$$L_{IOIM} = -\sum_{i=1}^{M} log\left(\frac{e^{\boldsymbol{v}_t^T \boldsymbol{x}_i / \tau_1}}{\sum_{j=1}^{C} e^{\boldsymbol{v}_j^T \boldsymbol{x}_i / \tau_1} + \sum_{k=1}^{Z} e^{\boldsymbol{u}_k^T \boldsymbol{x}_i / \tau_2}}\right). \quad (2)$$

In Eq. (2), the value of L_{IOIM} increases with τ_2, which makes the punishments between training identities and unlabeled identities further increscent. Fig. 2 shows the effects of the improved person search network with different loss functions. Comparing Fig. 2 (b) with Fig. 2 (a), we can see that IOIM loss makes different unlabeled identities much closer, while pushes the labeled identities and unlabeled identities further apart. Moreover, collecting lots of efficient samples from the same identity is unrealistic. It is difficult to distinguish thousands of identities when each identity has few samples. If we only use IOIM loss to learn the identification feature embeddings, the intra-class error is still very large. Considering the intra-class compactness of features learned by center loss [17], a multi-loss fusion strategy is proposed as follows:

$$L = L_{IOIM} + \frac{\lambda}{2} \sum_{i=1}^{M} \|\hat{\boldsymbol{x}_i} - \boldsymbol{d}_t\|_2^2, \quad (3)$$

where λ is the weight to balance two loss functions, $\boldsymbol{x}_i$ is the i-th unnormalized 256 dimensional feature embedding, and $\boldsymbol{d}_t$ represents the t-th class center of features. The first term, IOIM loss, is non-parametric and enables the model to converge much faster. The second term, center loss, is used to further minimize the intra-class distance by penalizing the distances between the features of samples and their centers. As depicted in Fig. 2 (c), the features of the same class are more centralized by using the proposed multi-loss fusion strategy.

2.4. Optimization

In order to enable the network to learn discriminative feature embeddings, we update the lookup table, the circular queue, and the class centers of different identities in backward computation. The t-th column of the lookup table $\boldsymbol{V}$ is updated in the (l+1)-th iteration by:

$$\boldsymbol{v}_t^{(l+1)} = \gamma \boldsymbol{v}_t^{(l)} + (1 - \gamma) \boldsymbol{x}_i, \quad (4)$$

where γ is the updating rate, and $\gamma \in [0, 1]$. In each iteration, the old terms of the circular queue are replaced with new features of unlabeled identities. In the (l+1)-th iteration, the center $\boldsymbol{d}_j$ of the j-th class in center loss is updated by:

$$\boldsymbol{d}_j^{(l+1)} = \boldsymbol{d}_j^{(l)} - \alpha \frac{\sum_{i=1}^{M} \delta(t = j) \cdot (\boldsymbol{d}_j^{(l)} - \hat{\boldsymbol{x}_i})}{1 + \sum_{i=1}^{M} \delta(t = j)}, \quad (5)$$

where α is the learning rate ranging from 0 to 1, and $\delta(\cdot)$ is the indicator function.

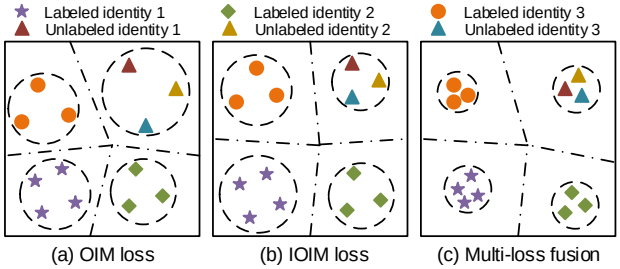

Fig. 2. The effects of the improved person search network with (a) OIM loss, (b) IOIM loss and (c) multi-loss fusion strategy in feature space (**Best viewed in color**).

3. EXPERIMENTS AND ANALYSIS

In this section, the proposed method is evaluated on benchmark CUHK-SYSU dataset [6] and PRW dataset [7]. We first describe the details of datasets and their evaluation setups, and then give our experimental results and analysis.

CUHK-SYSU dataset[1] contains 18,184 images, 8,432 labeled identities, and 99,809 annotated pedestrians bounding boxes in total. The training set contains 11,206 images and 5,532 identities, and the testing set contains 6,978 gallery frames and 2,900 query persons. Following [6], we use the protocol with gallery size equal to 100.

PRW dataset[2] contains 11,816 images, 933 labeled identities, and 34,304 annotated pedestrians bounding boxes. The training set contains 5,704 images and 483 identities, and the testing set contains 6,112 gallery frames and 2,057 query persons. Following [7], the gallery size is equal to 6,112.

Experimental settings. We implement our network based on Caffe [20]. The ResNet-50 [21] is chosen as the CNNs model, which is initialized by the ImageNet-pretrained model [21] in pretraining phase. The front 10 and following 6 residual units in CNNs model are respectively employed as the backbone network and identification network. The temperature parameters τ_1 and τ_2 of IOIM loss are set to 0.1 and 10, respectively. The parameters $\lambda = 0.01$, $\gamma = 0.5$, Z = 5000, D = 256 and $\alpha = 0.5$ are set for two datasets. We train the

[1] http://www.ee.cuhk.edu.hk/~xgwang/PS/dataset.html
[2] http://www.liangzheng.com.cn/Project/project_prw.html

Table 1. Performances of our IPSN using OIM loss, IOIM loss and multi-loss fusion strategy, and comparison with the baseline JDI+OIM on CUHK-SYSU and PRW datasets.

Method	CUHK-SYSU		PRW	
	mAP(%)	top-1(%)	mAP(%)	top-1(%)
JDI + OIM [1]	75.50	78.70	19.54	60.18
IPSN + OIM	78.81	79.45	19.96	61.84
IPSN + IOIM	79.15	79.55	20.35	58.43
IPSN + Multi-loss[1](Ours)	**79.78**	**79.90**	**21.00**	**63.10**

[1] Multi-loss means the proposed multi-loss fusion strategy.

Table 2. Comparisons among different end-to-end methods on the occlusion, low-resolution and whole testing set.

	E2E_PS [6]		JDI + OIM [1]		**Ours**	
	mAP(%)	top-1(%)	mAP(%)	top-1(%)	mAP(%)	top-1(%)
Occlusion	49.66	50.80	57.31	**58.82**	**60.51**	57.75
Low-res	49.95	53.45	55.56	59.66	**61.38**	**61.38**
Whole	69.69	72.97	75.50	78.70	**79.78**	**79.90**

model using Nesterov accelerated gradient decent [22] for two datasets with a NVIDIA GeForce GTX 1080 GPU. The learning rate is initially set as 0.001. All experiments are evaluated with mean Average Precision (mAP) and top-1 matching rate.

Analysis of the improved network and loss. Table 1 shows the results of improved person search network (IPSN) with OIM loss, IOIM loss, and multi-loss fusion strategy on CUHK-SYSU and PRW datasets. It can be observed that IPSN+OIM has better mAP and top-1 than joint detection and identification feature learning (JDI+OIM) [1] on two datasets. It is because that the pretraining phase in proposed IPSN provides proper initial state for the whole training network, and boosts the convergence of OIM loss. Moreover, IPSN+IOIM outperforms IPSN+OIM on CUHK-SYSU dataset, since the proposed IOIM loss makes the feature embeddings of unlabeled identities much closer by softening the probability distribution over unlabeled identities. IPSN+IOIM also has better mAP and competitive top-1 on PRW dataset. Furthermore, the proposed IPSN with multi-loss fusion strategy obtains the highest mAP and top-1 on two datasets, which shows the significance of penalizing the intra-class distances of all identities. The training loss curves of JDI+OIM and our method are shown in Fig.3. It can be seen that the loss curve of our method drops and converges much faster, which proves the effectiveness of the proposed method.

Evaluation on occlusion and low-resolution subsets. To evaluate the robustness of the proposed method, we conduct experiments on the occlusion and low-resolution subsets of CUHK-SYSU dataset. The results are shown in Table 2. It is observed that our method achieves better mAP and top-1 than other two end-to-end methods, namely end-to-end person search (E2E_PS) and JDI+OIM, on low-resolution subset. For occlusion subset, the results of our method is still competitive. Both occlusion and low-resolution subsets miss much appearance information of persons, so all the methods perform worse on these two testing subsets than on the whole testing set.

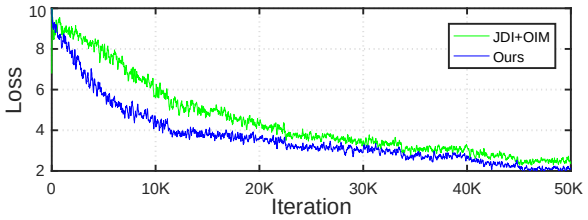

Fig. 3. The loss curves of JDI+OIM and the proposed method on CUHK-SYSU dataset (**Best viewed in color**).

Table 3. Compare proposed method with state-of-the-arts.

Method	CUHK-SYSU		PRW	
	mAP(%)	top-1(%)	mAP(%)	top-1(%)
ACF + LOMO_XQDA [1]	55.50	63.10	10.50	31.50
SSD + DLDP [23]	57.76	64.59	11.80	37.80
E2E_PS [6]	69.69	72.97	-	-
GT + DLDP [23] [1]	74.00	76.70	-	-
DPM + DLDP [23]	-	-	15.59	45.40
DPM_Alex + IDE$_{det}$ [7]	-	-	20.20	48.20
JDI + OIM [1]	75.50	78.70	19.54	60.18
Ours	**79.78**	**79.90**	**21.00**	**63.10**

[1] GT means the ground truth bounding boxes of pedestrians.

Comparision with state-of-the-arts. Table 3 shows the comparison results between the proposed method and state-of-the-arts on CUHK-SYSU and PRW datasets. Our method performs better than other combinations of pedestrian detectors and person re-identification algorithms, such as ACF+LOMO_XQDA [1], SSD+DLDP [23], DPM+DLDP [23], and DPM_Alex+IDE$_{det}$ [7]. It is because that our method utilizes a unified network to jointly learn feature embeddings for pedestrian detection and person re-identification, which can avoid some false detection alarms and misdetections. Moreover, our method achieves 79.78% mAP and 79.90% top-1 on CUHK-SYSU dataset, which outperform GT+DLDP [23] by 5.78% and 3.20%, respectively. It further implies that the end-to-end network can reduce the influence of the misdetections. While both E2E_PS [6] and JDI+OIM are end-to-end networks, our method significantly outperforms them. The results verify the learned feature embeddings in our method are more discriminative.

4. CONCLUSIONS

This paper presents an improved end-to-end network with multi-loss to learn discriminative feature embeddings for person search. The designed pretrained model provides proper initial state for the training network. Moreover, the proposed network reduces the pedestrian misdetections and false alarms by jointly optimizing pedestrian detection and identification. The proposed multi-loss fusion strategy makes full use of the information of unlabeled identities and makes the intra-class feature embeddings much closer. Experimental results on two benchmark datasets, CUHK-SYSU and PRW, demonstrate that our method achieves better mAP and top-1 than existing state-of-the-art methods. Results on two testing subsets confirm that our method is robust to occlusion and low-resolution.

5. REFERENCES

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

[2] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1995, 2009.

[3] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters (PRL)*, vol. 34, no. 1, pp. 3–19, 2013.

[4] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition (PR)*, vol. 68, pp. 346–362, 2017.

[5] Y. Xu, B. Ma, R. Huang, and L. Lin, "Person search in a scene by jointly modeling people commonness and person uniqueness," *in ACM International Conference on Multimedia (ACMM)*, pp. 937–940, 2014.

[6] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "End-to-end deep learning for person search," *arXiv preprint arXiv:1604.01850*, 2016.

[7] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian, "Person re-identification in the wild," *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1367–1376, 2017.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," *in European Conference on Computer Vision (ECCV)*, pp. 21–37, 2016.

[9] H. Liu and Q. Guan, "LPCV: Learning projections from corresponding views for person re-identification," *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1777–1781, 2017.

[10] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3908–3916, 2015.

[11] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition (PR)*, vol. 48, no. 10, pp. 2993–3003, 2015.

[12] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," *in International Conference on Machine Learning (ICML)*, pp. 507–516, 2016.

[14] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 212–220, 2017.

[15] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1249–1258, 2016.

[16] S. Wu, Y. Chen, X. Li, A. Wu, J. You, and W. Zheng, "An enhanced deep feature representation for person re-identification," *in IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–8, 2016.

[17] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," *in European Conference on Computer Vision (ECCV)*, pp. 499–515, 2016.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 6, pp. 1137–1149, 2017.

[19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *in Advances in Neural Information Processing Systems Workshops (NIPSW)*, 2014.

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *in ACM International Conference on Multimedia (ACMM)*, pp. 675–678, 2014.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[22] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," *in International Conference on Machine Learning (ICML)*, pp. 1139–1147, 2013.

[23] A. Schumann, S. Gong, and T. Schuchert, "Deep learning prototype domains for person re-identification," *in IEEE International Conference on Image Processing (ICIP)*, 2017.