INSTANCE ENHANCING LOSS: DEEP IDENTITY-SENSITIVE FEATURE EMBEDDING FOR PERSON SEARCH

Wei Shi[1], *Hong Liu*[1], *Fanyang Meng*[1,2], *Weipeng Huang*[1]

[1]Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, China
[2]Shenzhen Institute of Information Technology, China
{pkusw, hongliu, fymeng, wepon}@pku.edu.cn

ABSTRACT

Person search, which is vital for intelligent surveillance, aims at detecting and re-identifying pedestrians from whole monitoring images. However, due to the inaccurate pedestrian detections and extremely few instances per training identity, it remains challenging to learn discriminative representations only by labeled identities for person search. To this end, this paper proposes a novel loss function called instance enhancing loss (IEL) to learn deep identity-sensitive features by introducing unlabeled identity information. Specifically, the proposed IEL can selectively annotate unlabeled identities with similar appearances to labeled identities, and utilize these unlabeled identities in conjunction with labeled identities to train the person search network. The amount of unlabeled identities used as labeled instances can be quantitatively adjusted. Moreover, the proposed IEL is trainable and easy to optimize by back propagation algorithms. Extensive experiments on two benchmark datasets, namely CUHK-SYSU and PRW, show that our method outperforms state-of-the-arts for person search.

Index Terms— Person Search, Feature Embedding, Loss Function

1. INTRODUCTION

Person re-identification [1,2,3,4], a task of matching the query person across multiple non-overlapping camera views, is still challenging for real applications due to the heavy dependence on manually cropped bounding boxes. In fact, both automatic pedestrian detection and accurate person re-identification are necessary for real-world scenarios. For this reason, person search [5,6,7,8,9], which addresses pedestrian detection [10] and person re-identification simultaneously, has been a novel hot topic in intelligent surveillance and analysis [11,12,13].

Over the past few years, various methods for person search have been developed, which can be divided into two categories: indirect method and end-to-end method. Indirect methods [9] treat pedestrian detection and re-identification as two isolated parts, and sequentially combine them. However, inaccurate detections seriously affect the searching performance. To solve this problem, great attentions have been paid to end-to-end methods that treat pedestrian detection and re-identification as a joint optimization problem. Xu *et al*. [7] simultaneously utilized a Gaussian Mixture Model [14] to capture person commonness for detection, and applied Fisher vectors [15] to encode person uniqueness for identification. Xiao *et al*. [5] designed a unified Convolutional Neural Network (CNN) based model to learn both pedestrian detection and identification feature embedding. Zheng *et al*. [8] regarded person search as a fine-grained object detection [16] issue, and proposed a R-CNN [17] based model to address person search.

Although aforementioned end-to-end methods can reduce the influence of inaccurate detections, it is still very difficult to learn discriminative features for each identity. On one hand, there are many unlabeled identities in datasets, since manually annotating raw data will waste lots of labour. On the other hand, unlabeled identities cannot be directly used to learn feature embedding in a supervised training manner. To further enlarge the distances between labeled identities and unlabeled identities, Xiao *et al*. [5] proposed online instance matching (OIM) loss, which treats unlabeled identities as negatives for labeled identities in identification phase. Though the features learned by OIM loss have large inter-class variations, they lack of identity-sensitive property.

In this work, we observe that unlabeled identities with significant texture information can be used to enhance the labeled identities, so a novel instance enhancing loss is proposed to enhance identity-sensitive property of learned features. Firstly, to make full use of unlabeled identity information, all unlabeled identities are represented in the feature space learned by only labeled identities. Then, the distances between the deep feature of each unlabeled identity and feature centers of labeled identities are calculated for selectively annotating unlabeled identities. Finally, the annotated unlabeled identities as labeled instances are utilized in conjunction with labeled identities for discriminative feature learning. All of above processes are integrated into the proposed IEL to learn deep identity-sensitive features in an end-to-end manner. Furthermore, our IEL is easy to optimize, and can improve the robustness of learned features to inaccurate pedestrian detections.

This work is supported by National Natural Science Foundation of China (NSFC, No.U1613209), Scientific Research Project of Shenzhen City (No.JCYJ20170306164738129).

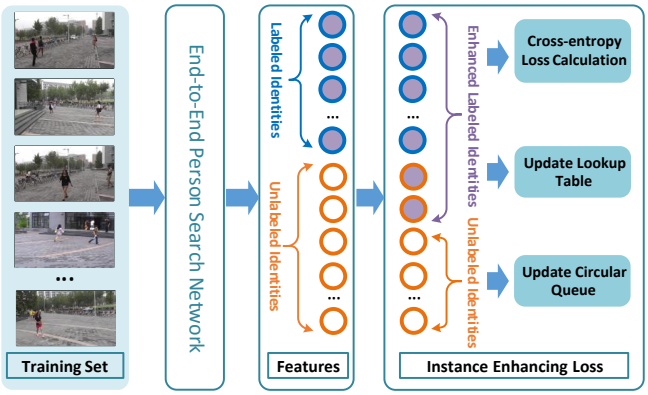


Fig. 1. The architecture of end-to-end person search network with proposed instance enhancing loss. The orange circles with blue padding represent the features of unlabeled identities that are annotated as labeled instances (**Best viewed in color**).

2. END-TO-END PERSON SEARCH NETWORK

The overall architecture of the end-to-end person search network is depicted in Fig. 1. The raw monitoring images captured from cameras contain not only pedestrians but complicated background as well. The person search network takes these images as training set to learn a joint model for both pedestrian detection and identification. Let $\mathbf{G}_k$ denote the k-th gallery image in training set and $\mathbf{P}_k^s$ stand for the s-th pedestrian in $\mathbf{G}_k$. Similar to the work [5], the person search network mainly consists of a backbone network, a region proposal network [18] and an identification network. The backbone network is utilized to learn common features for both pedestrian detection and identification. The obtained feature maps $\mathbf{F}_G^k$ are fed into the region proposal network to predict the locations of pedestrian candidates. According to the predicted locations, the feature maps of pedestrian candidates are cropped from $\mathbf{F}_G^k$, and then resized by RoI pooling layer [19]. The resized feature maps are sent to the identification network, generating learned feature embedding as follows:

$$\mathbf{X} = \left[\frac{\boldsymbol{x}_1}{\|\boldsymbol{x}_1\|}, \frac{\boldsymbol{x}_2}{\|\boldsymbol{x}_2\|}, \ldots, \frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|}, \ldots, \frac{\boldsymbol{x}_I}{\|\boldsymbol{x}_I\|} \right], \quad (1)$$

where $\boldsymbol{x}_i$ denotes the feature of the i-th pedestrian candidate in a batch, and $i \in [1, I]$. Divided by the L2-norm $\|\boldsymbol{x}_i\|$, the feature $\boldsymbol{x}_i$ is normalized to D-dimensional $\hat{\boldsymbol{x}}_i$ to restrict different identities to the same feature space. The normalized features are learned by the proposed instance enhancing loss. The implementation details of the network are shown in Section 4.

3. INSTANCE ENHANCING LOSS

For only dozens of instances per identity, it is difficult to train a fine-grained classification model with traditional loss functions, such as softmax loss [6]. This work presents a new loss function to integrate unlabeled identity information into person search. As shown in Fig. 1, the proposed instance enhancing loss (IEL) is connected with the person search network to learn deep identity-sensitive features in an end-to-end manner.

Fig. 2. Three groups of pedestrian instances from PRW dataset. For each group, "Ref" and "Pos" are two instances of the same labeled identity, and "Neg" is the unlabeled identity with high similarity to "Ref" (**Best viewed in color**).

3.1. Formulating the Instance Enhancing Loss

The online instance matching loss [5] is built on the softmax loss that contains a fully connected (FC) layer, a softmax layer and a cross-entropy loss layer. The OIM loss, which replaces the FC layer with a lookup table (LUT) $\boldsymbol{V} \in \mathbb{R}^{D \times C}$ and a circular queue (CQ) $\boldsymbol{U} \in \mathbb{R}^{D \times Z}$, can be written as:

$$L_{OIM} = -\sum_{i=1}^{M_1} log\left(\frac{e^{\boldsymbol{v}_{t(i)}^T \hat{\boldsymbol{x}}_i / \tau}}{\sum_{j=1}^{C} e^{\boldsymbol{v}_j^T \hat{\boldsymbol{x}}_i / \tau} + \sum_{k=1}^{Z} e^{\boldsymbol{u}_k^T \hat{\boldsymbol{x}}_i / \tau}} \right), \quad (2)$$

where M_1 is the number of labeled instances in current batch, and C is the number of all labeled identities. The term $\boldsymbol{v}_{t(i)}^T \hat{\boldsymbol{x}}_i$, in which $\boldsymbol{v}_{t(i)}$ is the $t(i)$-th column of $\boldsymbol{V}$, denotes the score of $\hat{\boldsymbol{x}}_i$ being classified as the $t(i)$-th labeled identity, and $\boldsymbol{u}_k^T \hat{\boldsymbol{x}}_i$ represents the similarity between $\hat{\boldsymbol{x}}_i$ and 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 lower temperature parameter leads to the harder probability distribution over different classes [20].

From Formula 2, we can see that the unlabeled identities have not been utilized to train the person search model, and all unlabeled identities are just treated as a negative class. However, as shown in Fig. 2, some unlabeled identities have very similar appearances to labeled identities, and they can be used to enhance the feature representations of labeled identities. To this end, we propose the instance enhancing loss as:

$$L_{IEL} = -\sum_{i=1}^{M} \lambda_i log\left(\frac{e^{\boldsymbol{v}_{|t(i)|}^T \hat{\boldsymbol{x}}_i / \tau}}{\sum_{j=1}^{C} e^{\boldsymbol{v}_j^T \hat{\boldsymbol{x}}_i / \tau} + \alpha \sum_{k=1}^{Z} e^{\boldsymbol{u}_k^T \hat{\boldsymbol{x}}_i / \tau}} \right), \quad (3)$$

where M is the number of all labeled and unlabeled samples in current batch, α is an adjustable parameter that is used to control whether the CQ is used or not. The training weight λ_i of the i-th training sample for loss calculation is defined as:

$$\lambda_i = \frac{1 + sign(t(i))}{2} + \frac{1 - sign(t(i))}{2} \cdot \frac{\eta}{1 + e^{-\gamma(d_i - \beta))}}, \quad (4)$$

with the maximal cosine distance

$$d_i = \max_{j \in \{1, \ldots, C\}} (\hat{\boldsymbol{x}}_i \cdot \boldsymbol{v}_j), \quad (5)$$

where η is the maximal training weight for unlabeled identities, and γ is the decay factor. The term d_i denotes the maximal cosine distance between the normalized feature $\hat{\boldsymbol{x}}_i$ of the unlabeled identity and the features of all labeled identities in LUT. Specifically, to distinguish from labeled identities, each unlabeled identity is annotated by the additive inverse of the label of the labeled identity that has maximal cosine distance d_i with the unlabeled identity. The parameter $\beta (-1 \leqslant \beta \leqslant 1)$ is the threshold that controls the number of unlabeled identities

treated as enhanced instances. The sign function $sign(t(i))$ = 1 if $0 < t(i) \leqslant C$, otherwise $sign(t(i))$ = -1.

As shown in Formula 3, the proposed IEL can selectively utilize these annotated unlabeled identities to enhance the corresponding labeled instances by the adjustable parameter β. Specially, when α and β are set to 1, the weights of all unlabeled identities in IEL will tend to zero. In this case, only labeled identities are treated as training samples, and the Formula 3 is exactly the OIM loss. Therefore, the proposed IEL is a generalized version of the OIM loss. When α and β are set to 0 and 1 respectively, the IEL only employs labeled identities to learn a base model, which is used to represent all unlabeled identities. To make the feature learning procedures with IEL clear, the detailed training steps are listed in Algorithm 1.

3.2. Optimization

The instance enhancing loss can be optimized by the back propagation (BP) algorithm, and it can pass on the losses by the chain rule. The gradients of IEL with respect to $\hat{\boldsymbol{x}}_i$ is:

$$\frac{\partial L_{IEL}}{\partial \hat{\boldsymbol{x}}_i} = \frac{\lambda_i}{\tau}\left(\sum_{j=1}^{C} p_j \boldsymbol{v}_j + \alpha \sum_{k=1}^{Z} q_k \boldsymbol{u}_k - \boldsymbol{v}_{|t(i)|}\right), \quad (6)$$

with

$$p_j = \frac{e^{\boldsymbol{v}_j^T \hat{\boldsymbol{x}}_i/\tau}}{\sum_{j=1}^{C} e^{\boldsymbol{v}_j^T \hat{\boldsymbol{x}}_i/\tau} + \alpha \sum_{k=1}^{Z} e^{\boldsymbol{u}_k^T \hat{\boldsymbol{x}}_i/\tau}}, \quad (7)$$

$$q_k = \frac{e^{\boldsymbol{u}_k^T \hat{\boldsymbol{x}}_i/\tau}}{\sum_{j=1}^{C} e^{\boldsymbol{v}_j^T \hat{\boldsymbol{x}}_i/\tau} + \alpha \sum_{k=1}^{Z} e^{\boldsymbol{u}_k^T \hat{\boldsymbol{x}}_i/\tau}}, \quad (8)$$

where p_j is the probability of $\hat{\boldsymbol{x}}_i$ being recognized as the j-th labeled identity, and q_k is the probability of $\hat{\boldsymbol{x}}_i$ being recognized as the k-th unlabeled identity. The proposed IEL uses non-parametric LUT and CQ to store features. In the (l+1)-th iteration, the $|t(i)|$-th column of LUT is updated by:

$$\boldsymbol{v}_{|t(i)|}^{(l+1)} = \delta \boldsymbol{v}_{|t(i)|}^{(l)} + (1-\delta)\hat{\boldsymbol{x}}_i, \quad (9)$$

where δ is updating rate. Although unlabeled identities can be utilized as enhanced labeled instances, they cannot be used to update LUT with larger δ, since they are not labeled identities after all. If the current training sample is an unlabeled identity with $d_i \geqslant \beta$, then it is used to update LUT with $\delta \in [0.5, 1]$. The updating rate δ is set to 0.5 for labeled identities. The CQ of fixed queue size Z is updated by constantly pushing new features of unlabeled identities and popping old ones.

Algorithm 1 Identity-sensitive features learning algorithm with proposed instance enhancing loss

Input: Training data $\{\mathbf{G}_k\}$, initialized parameters $\boldsymbol{W} \leftarrow \boldsymbol{W}_0$ of the network, initialized (LUT) $\boldsymbol{V}$ and (CQ) $\boldsymbol{U}$, parameters γ, η, δ, iteration number $l \leftarrow 0$ and learning rate.
Output: The parameters $\boldsymbol{W}$.

$\alpha \leftarrow 0$, $\beta \leftarrow 1$.
while not converge **do**
  $l \leftarrow l + 1$.
  Compute the IEL L_{IEL} by Formula 3.
  Compute the gradients $\frac{\partial L_{IEL}}{\partial \hat{\boldsymbol{x}}_i}$ by Formula 6.
  Update the parameters $\boldsymbol{W}$ by BP algorithm.
  Update $\boldsymbol{V}$ by Formula 9.
$\boldsymbol{W} \leftarrow \boldsymbol{W}_0$, $\alpha \leftarrow 1$, $l \leftarrow 0$, $\beta \leftarrow \beta^*$ ($\beta^* \in [0.5, 1]$)
Annotate the unlabeled identities according to Formula 5.
Compute the weights of all identities by Formula 4.
while not converge **do**
  $l \leftarrow l + 1$.
  Compute the IEL L_{IEL} by Formula 3.
  Compute the gradients $\frac{\partial L_{IEL}}{\partial \hat{\boldsymbol{x}}_i}$ by Formula 6.
  Update the parameters $\boldsymbol{W}$ by BP algorithm.
  Update $\boldsymbol{V}$ by Formula 9 and update $\boldsymbol{U}$.
return $\boldsymbol{W}$

4. EXPERIMENTS AND ANALYSIS

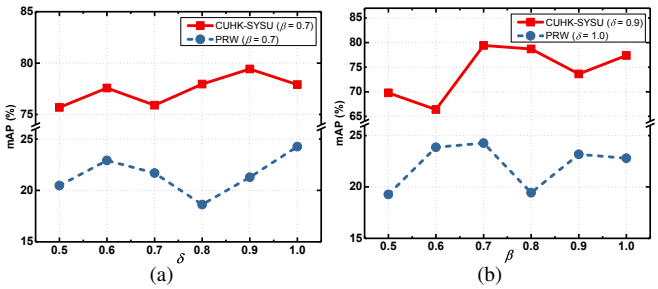
The proposed method is evaluated on two benchmark datasets: CUHK-SYSU[1] [6] and PRW[2] [8]. CUHK-SYSU dataset contains 11,206 images of 5,532 identities in training set, while PRW dataset contains 5,704 images of 483 identities for training. CUHK-SYSU dataset contains 2,900 query persons and a gallery of 6,978 images in testing set, and PRW dataset contains 2,057 query persons and a gallery of 6,112 images for testing. The identities are non-overlapping between training and testing set for both two datasets. Following [6], the gallery size is set to 100 if not specified for CUHK-SYSU, and the gallery size is set to 6,112 for PRW.

Implementation details: In the person search network, the backbone work is composed of a $64 \times 7 \times 7$ convolutional layer and the first 10 residual units in ResNet-50 [21]. The region proposal network [18] has the same setups as the work [5], which is utilized to predict the locations of pedestrian candidates. The identification network contains the last 6 residual units and an average pooling layer in ResNet50. Moreover, a 256-dimensional FC layer is used to represent the training samples. A two-dimensional FC layer with softmax loss, and an eight-dimensional FC layer with smoothed-L1 loss [19] are utilized to refine the results of pedestrian candidates.

Experimental settings: This work is implemented by Caffe [22] with a NVIDIA GeForce GTX 1080 GPU, and the network is initialized by the ImageNet pre-trained ResNet-50 model. Following [5], the queue size Z is set to 5000 for CUHK-SYSU, and the temperature τ is set to 0.1. Since there are only 483 training identities in PRW dataset, Z is set to 450 for balance. The parameters β and δ are evaluated in Fig. 2, and $\gamma = 20$, $\eta = 0.1$ are set for two datasets. In back propagation process, we train the network with Nesterov accelerated gradient decent [23], and the initial learning rate is set to 0.001. The mean Average Precision (mAP) and the top-1 matching rate are adopted as evaluation metrics.

Evaluation of parameters: Fig. 3 shows the sensitiveness of the proposed method with respect to the parameters, i.e. δ

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



(a) (b)

Fig. 3. Evaluation of the parameter δ and β on two datasets with one parameter changing and the other in default values.

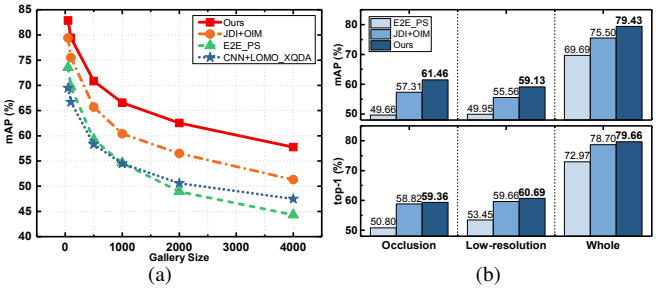
and β. If δ and β are less than 0.5, the unlabeled identities with low d_i will introduce many distractions to LUT update and identity-sensitive features learning, so both δ and β are evaluated from 0.5 to 1. In Fig. 3 (a), the mAP is the highest when $\delta = 0.9$ for CUHK-SYSU, and $\delta = 1.0$ for PRW. The results imply that the unlabeled identities cannot be utilized to update the LUT with big weights. In Fig. 3 (b), the performance is the best for two datasets when β is set to 0.7. If β is too small, more unlabeled identities with dissimilar appearances are used to train the model, which introduces many distractions. If β is too large, the unlabeled identities cannot play great role in enhancing the feature representations of labeled identities.

Evaluation of proposed loss: The effectiveness of the proposed IEL is evaluated as shown in Table 1. Similar to the baseline, IEL is also evaluated by training the end-to-end person search network (JDI) [5]. When $\alpha = 0$ and $\beta = 1$ are set for IEL, the JDI is learned only by training labeled identities. It can be seen that our method outperforms the other two approaches, because lots of unlabeled data are introduced as training samples to enhance the representations of labeled identities. Note that our method performs better than the baseline while using the ground truth (GT) boxes as a perfect detector. Moreover, comparing with the baseline, the improvement of our method under pedestrian detector is slightly higher than that under GT, since the identity-sensitive features learned by our method are robust to the inaccurate detections.

Table 1. Evaluation of our method on benchmark datasets.

Method	CUHK-SYSU		PRW	
	mAP(%)	top-1(%)	mAP(%)	top-1(%)
JDI + OIM (Baseline) [5]	75.50	78.70	21.52	66.55
JDI + IEL ($\alpha = 0$, $\beta = 1$)	77.39	77.69	22.79	69.03
JDI + IEL (**Ours**)	**79.43**	**79.66**	**24.26**	**69.47**
GT (Baseline)	77.90	80.50	25.22	71.12
GT (**Ours**)	**81.61**	**81.48**	**27.66**	**73.21**

Influence of various factors: Fig. 4 shows the performances of the proposed method and other methods on CUHK-SYSU dataset under different gallery sizes (Fig. 4 (a)) and extra factors (Fig. 4 (b)), i.e. occlusion and low-resolution. It can be seen that our method surpasses other methods by a notable margin under different gallery sizes. Person search mainly relies on the appearance information of persons, so its performance is heavily influenced by the occlusion and



(a) (b)

Fig. 4. Comparisons between the proposed method and other methods under different gallery sizes (a) and extra factors (b).

Table 2. Comparisons between our method and state-of-arts.

Method	CUHK-SYSU		PRW	
	mAP(%)	top-1(%)	mAP(%)	top-1(%)
ACF + LOMO_XQDA [5]	55.50	63.10	10.50	31.50
SSD + DLDP [9]	57.76	64.59	11.80	37.80
E2E_PS [6]	69.69	72.97	-	-
GT + DLDP [9]	74.00	76.70	-	-
DPM + DLDP [9]	-	-	15.59	45.40
DPM_Alex + IDE$_{det}$ [8]	-	-	20.20	48.20
JDI + OIM [5]	75.50	78.70	21.52	66.55
JDI + IEL (**Ours**)	**79.43**	**79.66**	**24.26**	**69.47**

low-resolution. However, our method still outperforms other methods on occlusion and low-resolution subsets, which implies our method is more robust to these extra factors, and can learn identity-sensitive features under inaccurate detections.

Comparisons with state-of-arts: Table 2 compares the proposed method with state-of-arts on CUHK-SYSU and PRW datasets. Our method performs better than some combinations of pedestrian detectors and person re-identification algorithms, such as ACF+LOMO_XQDA [5], SSD+DLDP [9], GT+DLDP [9], and DPM+DLDP [9]. Moreover, our method also outperforms some end-to-end methods, such as E2E_PS [6], DPM_Alex+IDE$_{det}$ [8] and JDI+OIM, since the proposed enhanced loss can integrate lots of unlabeled identity information to enhance the feature representations of labeled identities. Improvements of 3.93% mAP on CUHK-SYSU dataset and 2.74% mAP on PRW dataset are achieved over JDI+OIM, which can further verify the superiority of the proposed IEL.

5. CONCLUSIONS

In this work, we present a novel loss function called instance enhancing loss (IEL) to learn deep discriminative identity-sensitive feature embedding for person search. Specifically, the proposed IEL integrates the unlabeled identity information into feature learning process by selectively utilizing unlabeled identities as enhanced instances. Moreover, the proposed IEL is a generalized version of OIM loss, and easy to optimize by typical back propagation algorithms. Experimental results on benchmark CUHK-SYSU and PRW datasets show that our method achieves better mAP and top-1 than state-of-arts. Ablation studies on various factors verify the robustness of IEL against different gallery sizes, occlusion and low-resolution.

6. REFERENCES

[1] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 2, pp. 356–370, 2017.

[2] B. Mirmahboub, M. L. Mekhalfi, and V. Murino, "Person re-identification by order-induced metric fusion," *Neurocomputing*, vol. 275, pp. 667–676, 2018.

[3] H. Liu and W. Huang, "Body structure based triplet convolutional neural network for person re-identification," *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1772–1776, 2017.

[4] 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.

[5] 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.

[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] 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.

[8] 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.

[9] A. Schumann, S. Gong, and T. Schuchert, "Deep learning prototype domains for person re-identification," *in IEEE International Conference on Image Processing (ICIP)*, pp. 1767–1771, 2017.

[10] 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.

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

[12] 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.

[13] M. Liu, H. Liu, and C. Chen, "3D action recognition using multi-scale energy-based global ternary image," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2017.

[14] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 246–252, 1999.

[15] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," *in European Conference on Computer Vision (ECCV)*, pp. 413–422, 2012.

[16] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36, no. 8, pp. 1532–1545, 2014.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, 2014.

[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] R. Girshick, "Fast R-CNN," *in IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015.

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

[23] 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.