Anchor-Free Detector and Re-Identification with Joint Loss for Person Search

Ming Liu, Laifeng Hu, Yaonong Wang

ZheJiang Dahua Technology CO., LTD, Hangzhou, China

Abstract. Person search aims at matching a target person from a gallery of panorama images. Its performance depends on the localization accuracy and the recall rate of the pedestrian detector. FoveaBox detector which outperforms others is utilized in our framework. It generates high-quality region proposals for following re-identification (re-id). A joint loss function is proposed to train the network effectively. It is made up of on-the-fly Online Instance Matching (OIM) and proposal pair double margin contrastive (PPDMC) loss. We propose offset guided erasing instead of random erasing to solve the occlusion problem in person search task preferably. Experiments show that our method performs more effectively against the state-of-the-art methods on two widely used person search datasets.

1. Introduction

Person re-id [1] is a fundamental yet challenging task in computer vision which matches a target person from a gallery of images. It is used in real world video surveillance applications, such as looking for lost people, cross-camera tracking, etc. The task is challenging due to complex variations of human poses, lighting, occlusion, resolution. Therefore the task draws much attention in recent years [13], [23].

Traditional person re-id tasks contain two sub-problems, detecting all persons in the whole image and cropping them to extract features for identification. There is still a gap between its defects and practical applications. To bridge this gap, some recent works [3] propose the person search task that joins the task of detection with re-id by an end-to-end manner. Recently, several state-of-the-art methods [16], [24] have been proposed to improve the performance. Nevertheless, most of them focus on an end-to-end framework based on Faster R-CNN which is an anchor-based detector. The anchor-based detector has some drawbacks. First, anchor boxes import additional hyper parameters. Second, one kind of designed anchors based on a particular dataset is not always suitable to other datasets, leading to poor versatility. Third, a large set of anchor boxes are densely placed on the input image, resulting in a huge imbalance between positive and negative anchor boxes, and involving complicated and time-consuming computation. Therefore, some works trying to discard anchor boxes such as FoveaBox [18], FCOS [15] are proposed recently. FoveaBox detector is utilized in our framework in order to make an improvement on recall rate and detection accuracy.

OIM proposed by [3] is a method learning features to distinguish between different identities. As it uses fixed parameters to update features in a lookup table (LUT), the parameters need to be carefully designed. To make it more self-adaptive, we propose to update features with intersection-over-union (IOU).

OIM stores the features of labeled identities in LUT and unlabeled identities in Queue, and uses SoftMax [10] to do classification. Because there are a large number of identities, while each identity
only has several instances, it is intractable to learn features to distinguish between similar persons. Meanwhile, OIM loss only considers labeled and unlabeled identities, while the background clusters are left out. Therefore, background clusters which are similar to the labeled identities cannot be distinguished effectively. Inspired by the above discussion and given that it is impossible to have the same person in one image, we adopt proposal pairs in a Siamese architecture and join double margin contrastive (DMC) loss [7] with OIM loss to supervise the whole network.

Occlusion is a major challenge in person search task. To overcome the problem, a data augmentation, random erasing is proposed in [14]. In the training phase, we apply offset guided erasing instead of random erasing to proposals produced by the detector, which is more effective.

In summary, there are several contributions of our paper. First, we utilize anchor-free instead of anchor-based detector. Then we adopt proposal pairs in a Siamese architecture and join DMC loss with the on-the-fly OIM loss to supervise the whole network. Finally, we propose offset guided erasing which improves the recognition performance and robustness of re-id of partially occluded samples. The proposed method achieves 89.4% and 34.3% in mAP on CUHK-SYSU [3] and PRW [2] respectively, which outperforms state-of-the-art methods.

2. Related Work

Pedestrian Detection. Recent years, with the development of CNN, CNN-based methods have been developed, which can be roughly divided into the anchor-based [10], [20] and anchor-free [15], [18] manners. The main difference between anchor-based and anchor-free is whether utilization of anchor boxes which should be carefully designed based on training/validation set. Lin et al. [21] design Feature Pyramid Networks (FPN) architecture with lateral connections for building multi-level semantic feature maps to detect objects at different scales. The pedestrian detector in our approach is based on FoveaBox and FPN, which is also proved to be effective and reliable in the following experiments.

Person Re-Identification. Traditional person re-identification methods addressed the problem by hand-crafted discriminative features [26] and learning distance metrics [27]. Recently deep learning is introduced into the re-ID research to learn the feature for the following distance metrics. Li et al. [28] proposed a method that input a pair of cropped pedestrian images and utilize a binary verification loss function to train the model. Sun et al. [29] propose the Part-based Convolutional Baseline (PCB) to extract several body parts features. Zhun et al. [14] propose random erasing to yields consistent improvement over strong baselines in person re-identification. In this paper, we propose offset guide erasing to overcome the shortcomings of random erasing.

Person Search. To facilitate real-world person re-id, person search is developed which aims at matching a target person among the gallery persons in the whole images [3]. Most works [16], [24] train the detection and re-ID model in an end-to-end framework. Xiao et al. [3] employ the ResNet-50 as backbone network, and share the base layers with detector and identification. Meanwhile, they proposed Online Instance Matching (OIM) which outperforms in the classification task with large but sparse identities. In our method, we propose an end-to-end framework to learn the detector and re-identification for person searching, which joining OIM loss and proposal pair double margin contrastive (PPDMC) loss to supervise the whole network.
3. The Proposed Method

Real person search scenarios have two steps. The location of the target is detected in a whole image and then the satisfied boxes are sent into the re-id network. So we propose person search term consists of two stages, pedestrian detection and re-id which are shown in Fig. 1. We use anchor-free detector instead of traditional anchor-based detector in our framework. We choose FoveaBox detector with a panoramic image as input, use stem CNN to transform raw pixels to convolutional feature maps, and build a detection net consisting of two task-specific sub-networks upon these feature maps. We get a large number of coarse boxes along with confidence from per pixel classification and per pixel bounding box prediction subnet. Only prediction boxes which satisfy some corresponding regulations are fed into the identification net. Each of them uses RoiPooling to get a fixed dimensional output (256-D). This output vector is normalized by its L2-norm. In inference stage, we rank the gallery people according to their Euclidean distance to the query target person. In training stage, we propose DMC loss for proposal pairs to join with modified OIM loss on the top of features to supervise the identification net, together with focal loss for training the classifier and Smooth-L1 loss [10] for training the box regression in a multi-task framework. Our method is elaborated in detail in the following section.

3.1. Pedestrian Detection and Feature Learning

We use FoveaBox detector which is completely anchor-free framework that does not rely on the default anchors both in training and inference phase. It is more robust and denser to predict bounding box distributions on each position and mitigate the foreground-background classification and box regression challenge contrast to anchor-based detection framework such as Faster-RCNN [10].

The FoveaBox architecture is composed of a base network for feature extraction and two task-specific subnetworks for classification and location prediction. In our paper, we employ Feature Pyramid Network [21] based on ResNet-50 as the base network. We construct a pyramid with levels \( \{P_l\}, l=3,4,5 \), and each level of the pyramid feature maps built by top-down architecture with lateral connection can be used to detect objects at a different scale \( P_1 \) is \( 1/2^l \) of the input in resolution. Each pyramid level produces 256 channel feature maps and attaches two sub-networks, one for classifying the corresponding cells and the other for predicting the coordinates \((x_1,y_1,x_2,y_2)\) of object boxes in the scale range. Classification and box regression are based on per pixel with focal loss for training the category and Smooth-L1 loss for training the location. In inference, it produces a large number of boxes with confidence scores. First, we use a threshold \( \theta \) to screen out boxes with low confidence. Then, we select the top \( N_{box} \) from each prediction layer to send into non-maximum suppression (NMS) [10] with threshold \( \theta_{NMS} \) which is applied for each class separately. Finally, we select the top \( N_{per} \) boxes for each image. All candidate proposals selected are fed into RoIPooling layer to generate standard scale feature representations. Then attach identification net with average pooling layer to generate 2048-D vectors. Finally, the output is attached with fully connected layer and followed by L2-Norm for transforming feature where cosine similarities with the target person are computed when doing inference. During the training stage, we supervise the feature learning with our proposed method that joins PPDMC loss with on-the-fly OIM loss function.

3.2. On-the-fly Online Instance Matching Loss

To learn a robust and discriminative feature, the loss function needs to be designed carefully. Although Softmax loss is widely used in classification tasks, there are two drawbacks. First, large-scale person search datasets have a large number of identities and each identity has unbalanced instances or each image only contains few identities. Second, there are some non-specific class-ids and unlabeled identities in data-sets. It is intractable to exploit with SoftMax loss. Xiao et al [3] propose OIM loss to minimize the features discrepancy among the instances of the same person, while maximize the discrepancy among different people. In our work, we modify OIM loss to learn a more discriminative feature.
Fig. 2: The left part shows the labeled (blue) and unlabeled (orange) identity proposals in an image. The up-right part shows update LUT according to the id. The down-right part shows the valuable information of proposal increase with IOU increase.

As shown in Fig. 2, OIM maintains a LUT $V \in \mathbb{R}^{D \times L}$ to store the features of all the labeled identities. Denote $x \in \mathbb{R}^D$ as the feature of a labeled identities in mini-batch, where $D$ is the dimension of feature and $L$ is the number of different labeled identities. During the forward propagation, they compute cosine similarities between the mini-batch samples and the labeled identities within LUT as $V^T x$. Accordingly, during the backward propagation, they will update the $t$-th column of the LUT if the class-id of batch sample $x$ is $t$, by $v_t \leftarrow v_t + (1 - \gamma)\chi$, where $\gamma \in [0, 1]$, then scale $v_t$ normalized by $l_2$-Norm. In our approach, we import the IOU between detection boxes and ground-truth (gt) to replace constant momentum $\gamma$.

In training phase, the predicted boxes which are produced by the branch of box regression sub-network are not accurate and contain different target information. The valuable information of proposal increases with IOU growth as it is shown in the bottom-right part of . Therefore it is not reasonable to update the LUT $v_t$ with a constant parameter. The proposed technique is elaborated as below.

$$v_t \leftarrow (1 - \gamma)\left(1 - \frac{|iou - T|}{1 - T}\right)v_t + \gamma\frac{|iou - T|}{1 - T} x$$

where $T$ is the threshold of IOU in the training phase.

### 3.3. Proposal Pair Double Margin Contrastive Loss

There are three different types of proposals, labeled identities, unlabeled identities and background clusters. The OIM uses the LUT to store the features of all the labeled identities. Similar persons cannot be distinguished exactly from great variation across different scenes. An example is illustrated in Fig. 2 in which the left four women in blue bounding boxes are easy to confuse. Due to false positive from pedestrian det-ec tion and because OIM left out background clusters similar to labeled identities, background clusters cannot be disting- uished. Based on these two considerations, we propose PPDMC loss to assist OIM loss for learning more robust and discriminative features. Siamese-Net has two input fields and one output whose state value corresponds to the similarity between the two patterns. LeCun et al [5] propose to train a network consisting of two identical convolutional networks which share weights for learning a similarity metric. Then minimize a discriminative loss function that drives the similarity metric to be small for pairs of faces from the same person, and large for pairs from different persons. As shown in Fig. 3, we extend it to proposal produced by the detector to make up many positive and negative proposal pairs for learning a similarity metric. Positive proposal pair is the same identity positive proposal denoted as $PPP \in \{(p_i, p_j)\}$, where $id(p_i) = id(p_j)$, $p_i, p_j \in P$. Denote positive proposal as $P \in \{IOU(pred_{box}, gt) \geq TP_{thresh}\}$. Negative proposal pairs contain two types of pairs, different identity positive pairs and positive and negative pairs. Negative proposal is denoted as $N \in \{IOU(pred_{box}, gt) < FP_{thresh}\}$. Denote negative proposal pairs as $NNP$, where
Fig. 3. Architecture of the positive and negative proposal pairs.

\[ NPP = \begin{cases} (p_i, p_j), & p_i, p_j \in P, \text{id}(p_i) \neq \text{id}(p_j) \\ (p_i, n_j), & p_i \in P, n_j \in N \end{cases} \]  \hspace{1cm} (2)\]

Contrastive loss is a single margin method known for its simplicity and outperformance, but it is weak to process unbalanced positive and negative image pairs. In our work, proposal pairs \((PPP, NNP)\) utilized during training phase is seriously imbalanced. Hao et al [7] propose DMC loss which is especially suitable to solve unbalanced the problem which is formulated as,

\[ L(p_i, p_j) = \frac{1}{2} \left[ (y^* \max(\alpha, 0))^2 + (1-y^*) \max(\beta - d, 0)^2 \right] \]  \hspace{1cm} (3)\]

where \(d\) is the Euclidean distance, denoted as \(d = \|f(p_i)-f(p_j)\|\) and \(\alpha, \beta\) are the margin enforced for positive and negative proposal pairs, respectively.

3.4. Offset Guided Erasing Data Augmentation

Random erasing is a data augmentation method on images for training the convolutional neural network which reduces the risk of over-fitting and makes the model robust to occlusion in person re-id. As shown in Fig. 4, the proposals fed into re-id are random located around the gt. It is likely to erase useful information. To simulate more realistic occlusion

Fig. 4: Examples of Random Erasing for object detection. (b) Examples of offset guide erasing for object detection. (c) Different type of erasing exhibition on the proposal.

We propose offset which is produced by the center point of predicted proposal and gt difference value to guide erasing, which is formed as \(O_i = (O_{ci}, O_{ci}),\) where \(O_{ci} = (c_{x_{pred}} - c_{x_p}, c_{y_{pred}} - c_{y_p}), (c_{x_{pred}} - c_{x_{pred}}', c_{y_{pred}} - c_{y_{pred}}')\) and \((c_{x_p}, c_{y_p})\) are the center of proposal and gt respectively. If \(O_{ci} \leq 0\) and \(O_{ci} \leq 0\) the predicted box is placed on the top-left corner of gt, and we randomly erase on upper-left quarter of proposal. The proposed augmentation not only reserves useful foreground information, but also inhibits background information which improves performance in comparison with random erasing.

4. Experimental Results

In this section, we first introduce two common person search datasets and evaluation protocols. Then we present details of implementation and contributions of different constituent part aforementioned. Finally, we show exper-imental results in comparison to state-of-the-art methods.
4.1. Datasets and Evaluation Protocols

CUHK-SYSU. CUHK-SYSU consists of street/urban scene images captured by a hand-held camera or selected from movie snapshots. It contains 18184 images and 96143 pedestrian bounding boxes with 8432 labeled identities and 87711 unlabeled identities. We adopt the standard train/test split provided by the dataset, where the training set includes 11206 images and 5532 identities, and the testing set contains 2900 probe persons and 6978 gallery images. Moreover, each probe person corresponds to several gallery subsets with different sizes, which are defined in the dataset.

PRW. PRW is extracted from video frames captured with six cameras in a university campus. There are 11816 frames annotated with 34304 bounding boxes. Among all the pedestrians, 932 identities are tagged and the rest are marked as unknown. The training set includes 5134 images and 482 identities. The testing set contains 2057 probe persons and 6112 gallery images. Different from CUHK-SYSU, each probe person corresponds to the whole gallery subsets.

Evaluation Protocols. We adopt two evaluation metrics, the cumulative matching curve (CMC top-k) and the mean Average Precision (mAP) which is mentioned in [3] as our experiment performance measure. The top-k is widely used as performance metrics in re-id task. The mAP in re-id is based on the precision-recall curve in the object detection criterion which reflects mean accuracy and matching rate of searching a probe person from gallery images.

4.2. Implementation Details

We use an implementation of FoveaBox in open source object detection toolbox based on mmmdetection which is developed by Multimedia Laboratory, CUHK. The backbone network is ResNet-50 with FPN. Our detector model is pretrained on the SYSU and PRW datasets and used to initialize the entire framework model. Pair of images with persons of the same identity which are sized to $800 \times 600$ are taken as a mini-batch to be sent to stem CNN. We join focal loss, smooth-11 loss, OIM loss with PPDMC loss to supervise the whole network. We obtain 1000 coarse proposals by two sub-networks and select 128 proposals containing all positive proposals and several negative proposals to feed into the re-id network to learn features. Selected proposals are made pairs as aforementioned, where $TP_{thresh} = 0.5$ and $FP_{thresh} = 0.4$. Then we put proposal pairs into RoiPooling layer, offset guided erasing and identification net in turn and extract 256-D L2-norm features. Finally, we feed features of 128 batches into the on-the-fly OIM loss where hyperparameter $\ell = 0.5$ and feed 64 pairs of features into DMC loss where $\alpha = 0.8$ and $\beta = 1.2$. Then we assign different loss weights to the OIM loss and PPDMC loss. OIM loss weighted 0.9 while PPDMC loss weighted 0.1. We choose the batched Stochastic Gradient Descent (SGD) optimizer with the momentum of 0.9. As for the learning rate strategy, we use the warm-up strategy. The initial learning rate is 0.001, which warms up to 0.005 in the first 500 iterations, then decays to $1 \times 10^{-4}$ after 8 epochs and $1 \times 10^{-5}$ after 11 epochs. Our model is trained for 12 epochs in total. All experiments are conducted on PyTorch and the network is trained on 4 NVIDIA TI 1080i GPUs.

4.3. Comparison with the State-of-the-art Methods

In this subsection, we compare with several state-of-the-art methods on the CUHK-SYSU and PRW datasets with metrics of top-1 and mAP. Joint detection and search methods such as IAN [6], NPSM [8], RCAA [19], I-Net [4], MGTS [22], DisGCN [13], GCN [24] and CLSA [17] are analyzed. In addition, we also compare with some methods which separate person search in several steps including pedestrian detection (CNN [10]) and person descriptors (DSIFT [9]) and distance metrics (XQDA[12]).

Results On CUHK-SYSU. The results of ours and other state-of-the-art methods are listed in TABLE I. The CNN is the shortened form of Faster RCNN detector based on ResNet50 and IDNet represents re-id net. Compared with methods breaking down the problem into separate detection and re-id task, there is a significantly improvement of pro- posal person search framework. OIM is the baseline of the overall proposed person search framework. I-Net [4] intro- duces a Siamese structure and joins online pairing loss with hard example priority Softmax loss to learn robust feature representation. I-Net achieves about 1% improvement on both top-1 and mAP. MGTS [22] is a two-
team framework that adds segmentation network to generate clean foreground objects for feature learning. This method achieves great performance. GCN [24] designs contextual graph representation and contextual instance structure to improve performance. Compared with the state-of-the-art method CLSA [17], our method achieves 3.2%/2.2% gain on top-1 and mAP metrics.

Table 1 results on cuhk-sysu with agallery size of 100

| Method                  | top-1(%) | mAP (%) |
|-------------------------|----------|---------|
| CNN +DSIFT + Euclidean   | 39.4     | 34.5    |
| CNN + BOW[11] + Cosine   | 62.3     | 56.9    |
| CNN + LOMO[12]+XQDA     | 74.1     | 68.9    |
| CNN + IDNet             | 74.8     | 68.6    |
| OIM                     | 78.7     | 75.5    |
| IAN                     | 80.1     | 76.3    |
| NPSM                    | 81.2     | 77.9    |
| RCAA                    | 81.3     | 79.3    |
| I-Net                   | 81.5     | 79.5    |
| MGTS                    | 83.7     | 83.0    |
| DisGCN                  | 83.4     | 81.3    |
| GCN                     | 86.5     | 84.1    |
| CLSA                    | 88.5     | 87.2    |
| Ours                    | 91.7     | 89.4    |

Table 2 results on prw

| methods                  | top-1 (%) | mAP (%) |
|--------------------------|-----------|---------|
| LDCF+LOMO+XQDA           | 31.1      | 11.0    |
| LDCF[25] + IDE           | 44.6      | 18.3    |
| LDCF+IDE + CWS[2]        | 45.5      | 18.3    |
| OIM                      | 49.9      | 21.3    |
| NPSM                     | 53.1      | 24.2    |
| CLSA                     | 65.0      | 38.7    |
| DisGCN                   | 69.8      | 29.5    |
| MGTS                     | 72.1      | 32.6    |
| GCN                      | 73.6      | 33.4    |
| Ours                     | 74.1      | 34.3    |

Table 3 comparison with different detection

|                | All        | Labeled   | top-1  | mAP  |
|----------------|------------|-----------|--------|------|
|                | Recall     | AP        |        |      |
| RCNN+ours      | 83.64      | 78.13     | 98.68  | 88.59|
| RetinaNet+ours | 82.62      | 77.02     | 98.51  | 87.48|
| FoveaBox+ours  | 91.83      | 82.24     | 99.58  | 91.7 | 89.4 |

In order to verify the scalability performance of our method, we compare with other person search methods with different gallery sizes [50, 100, 500, 1000, 2000, 4000] as is shown in Fig. 5. As can be seen from the figure, there are different performance degenerations as the gallery size increases in all methods except CLSA. However, ours puts up with great advantages under different gallery sizes, which indicates the robustness of our method. Besides, we notice that our method is slightly worse than CLAS when the gallery size increases from 500 to 4000, while vastly better than CLAS when the gallery size is under 500.

Results On PRW. We report the results of our method comparing with other state-of-the-art methods on the PRW which are shown in TABLE III. Compared with CUHK-SYSU, all methods mentioned achieve poorer results in terms of top-1 and mAP metrics. This is mainly because the training set is less and the gallery size is the number of all the testing images. On PRW dataset, our method achieves 74.1% in top-1 accuracy and 34.3% in mAP which outperforms previous state-of-the-art methods. It proves the effectiveness of our method in the person search task.
Table 4 comparison with different loss types

| Loss type                      | top-1 (%) | mAP (%) |
|-------------------------------|-----------|---------|
| OIM                           | 87.5      | 84.9    |
| OIM with IOU                  | 88.7      | 86.2    |
| OIM with PPDMC                | 90.8      | 88.2    |
| OIM with IOU and PPDMC        | 91.7      | 89.4    |

TABLE 5 COMPARISON WITH DIFFERENT DATA AUGMENTATIONS

| augment                      | top-1 (%) | mAP (%) |
|-------------------------------|-----------|---------|
| baseline                     | 88.6      | 86.5    |
| Random erasing               | 90.3      | 88.1    |
| Offset guide erasing         | 91.7      | 89.4    |

4.4. Ablation Study

In this section, we first show a performance improvement of anchor-free detector comparing with anchor-based with the refinement of our methods. Second, we show the effectiveness of joining on-the-fly OIM loss with PPDMC loss. Finally, we show the influence of random and offset guided erasing on re-id. The experiments are conducted and analyzed on the CUHK-SYSU at a gallery size of 100.

FoveaBox Detector. As is known that in the field of pedestrian search, a good detector is very important to person re-id which needs a high recall rate of pedestrian detection and accurate localization. We compare different detectors including Faster R-CNN, RetinaNet [20] and FoveaBox. As the results shown in TABLE IV, the method of FoveaBox together with re-id is effective and obviously improves in both detector and re-id evaluation metrics.

Analysis of Loss. First, we propose on-the-fly OIM which considers the IOU factor. OIM uses fixed parameters to update features in LUT, while our method updates features adaptively which contains more foreground information in training phase. Then we join PPDMC loss with OIM loss to mutually supervise feature learning. To have an insight of the co-training between OIM loss and PPDMC loss, four different cases have been discussed, OIM only, OIM with IOU, OIM with PPDMC and OIM with IOU and PPDMC. TABLE V shows the results of 4 cases of loss types. We can see that single OIM loss achieves the worst result. By replacing fixed parameters with IOU to update features in the LUT, the mAP and top-1 both increase significantly. Further, the PPDMC loss can be considered as a special supplement of OIM loss, and the result of the joint loss of OIM and PPDMC increases sharply. Finally, joining OIM with IOU with PPDMC shows the best re-id result.

Offset Guided Erasing. Data augmentations such as random cropping, flipping and rotating are usually used in image classification, object detection and person re-id. Random erasing is more commonly applied in re-id and improves the result of recognition obviously. Considering the particularity of our framework, we propose offset guided erasing. The results of different types of
erasing are shown in TABLE VI. We can see that there is an improvement in top-1 and mAP in comparison with the baseline. This confirms the effectiveness of our method.

5. Conclusion
In this work, we propose an end-to-end framework for person search based on FoveaBox detector which outperforms in recall and precision rate. Experimental results demonstrate that our method of joining on-the-fly OIM loss and PPDMC loss to supervise the whole network is effective and useful. Finally, we propose offset guided erasing in the training phase of re-id and our method is more effective and robust in comparison with no erasing and random erasing. Our proposed framework achieves state-of-the-art performance on two widely adopted person search datasets, CUHK-SYSU and PRW.

References
[1] N. Gheissari, B. Sebastian, and R. Hartley. “Person re-identification using spatiotemporal appearance.” In CVPR, 2006.1.
[2] Liang Zheng, Hengsheng Zhang, and Shaoyan Sun. “Person re-identification in the wild.” In CVPR, pp. 3346–3355, 2017.
[3] Tong Xiao, Bochao Wang, Liang Lin, and Xiaogang Wang. “Joint detection and identification feature learning for person search.” In CVPR, pp. 3376–3385, 2017.
[4] Zhenwei He, Lei Zhang, and Wei Jia. “End-to-end detection and re-identification integrated net for person search.” In CoRR, 1804.00376, 2018.
[5] S. Chopra, R. Hadsell and Y. LeCun. “Learning a similarity metric discriminatively, with application to face verification.” In CVPR, pp.539–546, 2005.
[6] Jingxiang Xiao, Yanchun Xie, Kaizhu Huang, and Jiashui Feng. “IAN: the individual aggregation network for person search.” In Pattern Recognition, 87:332–340, 2019.
[7] Jiedong Hao, Jingdong Wang, Wei Wang, and Tieniu Tan. “DeepFirearm: Learning Discriminative Feature Representation for Fine grained Firearm Retrieval.” arXiv preprint arXiv: 1806.02984v2(2018).
[8] Hao Liu, Jiashi Feng, Jayashree K., and Shuicheng Yan. “Neural person search machines.” In ICCV, pp. 493–501, 2017.
[9] Martin Kostinger, Martin Hirzer, Paul Wohlhart, and Horst Bischof. “Large scale metric learning from equivalence constraints.” In CVPR, pp. 2288–2295, 2012.
[10] R. Girshick. “Fast r-cnn.” In Proceedings of the IEEE international conference on computer vision, pp. 1440–1448, 2015.
[11] L. Zheng, L. Shen, and Qi T. “Scalable person re-identification: A benchmark.” In ICCV, pp. 1116–1124, 2015.
[12] Shengcai Liao, X. Zhu, and Stan Z. Li. “Person re-identification by local maximal occurrence representation and metric learning.” In CVPR, pp. 2197–2206, 2015.
[13] Sofia IraKtena, Sarah Parisot, B. Glocker, and Daniel Rueckert. ”Distance metric learning using graph convolutional networks: Application to functional brain networks.” In MICCAI, pp.469–477, 2017, 14.
[14] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. “Random erasing data augmentation.” arXiv preprint arXiv: 1708.04896(2017).
[15] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. “FCOS: Fully Convolutional One-Stage Object Detection.” arXiv preprint arXiv: 1904.01355v5(2019).
[16] Chuchu Han, Jiacheng Ye and Yunshan Zhong. “Re-ID Driven Local-ization Refinement for Person Search.” arXiv preprint arXiv: 1909.08580 v1(2019).
[17] Xu Lan, Xiataian Zhu, and Shaogang Gong. “Person search by multi-scale matching.” In ECCV, 2018.
[18] Tao Kong, Huaping Liu, Yuning jiang, and Jianbo Shi. “FoveaBox: Beyond Anchor-based Object Detector.” arXiv preprint arXiv: 1904.03 797v1(2019).
[19] X. Chang, Y. Yang, and A. G. Hauptmann. “Rcaa: Relational context-aware agents for person search.” In ECCV, 2018.
[20] T.Y. Lin, R. Girshick, K. He and P. Dollar. “Focal loss for dense object detection.” In PAMI, 2018. 1, 2.
[21] P. Dollar, Girshick, K. He, and Belongie. “Feature pyramid networks for object detection.” In CVPR, vol.1, pp. 4, 2017.
[22] Di Chen, Wanli Ouyang, and Ying Tai. “Person search via a mask-guided two-stream CNN model.” In ECCV, pp 764–781, 2018.
[23] W. Li, R. Zhao, T. Xiao, and X. Wang. “Deepreid: Deep filter pairing neural network for person re-identification.” In CVPR, 2014. 1, 2, 7.
[24] Y. Yan, Q. Zhang, B. Ni, Wendong Zhang, and Ming hao Xu. “Learning Context Graph for Person Search.” arXiv preprint arXiv: 1904. 01830v1 (2019).
[25] W. Nam, Piotr Dollar, and J. H. Han. “Local decorrelation for improved pedestrian detection.” In NIPS, 2014.
[26] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. “Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences.” In ICDSC, 2008. 2.
[27] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. “Person re-identification by support vector ranking.” In BMVC, 2010. 2.
[28] W. Li, R. Zhao, T. Xiao, and X. Wang. “Deepreid: Deep filter pairing neural network for person re-identification.” In CVPR, 2014. 1, 2, 7.
[29] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. “Beyond part models: Person retrieval with refined part pooling.” In European Conference on Computer Vision, 2018.