EURASIP Journal on Wireless Communications and Networking

RESEARCH

Open Access

ABOS: an attention-based one-stage framework for person search

Yuqi Chen[1], Dezhi Han[1*], Mingming Cui[1], Zhongdai Wu[2,3] and Chin-Chen Chang[4]

*Correspondence:
dezhihan88@sina.com

[1] Department of Engineering, Shanghai Maritime University, Shanghai, China
[2] Shanghai Ship and Shipping Research Institute Co., Ltd., Shanghai, China
[3] COSCO Shipping Technology Co., Ltd., Shanghai, China
[4] Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan

Abstract

Person search is of great significance to public safety research, such as crime surveillance, video surveillance and security. Person search is a method of locating and identifying the queried person from a complete set of images. The main cause of false recall and missed detection in person search is the presence of person occlusion in the images. In order to improve the accuracy of person search when the person to be queried is occluded, this paper proposes an attention-based one-stage framework for person search (ABOS) using an anchor-free model as a baseline. The method uses the channel attention module to express different forms of occlusion and take full advantage of the spatial attention module to highlight the target region of the occluded pedestrians. These attention modules integrate deep and shallow features to guide the network to pay attention to the visible area of the occluded target and extract the semantic information of the pedestrians. Experimental results on CUHK-SYSU and PRW datasets show that the proposed person search method based on attention mechanism in this paper has better performance than existing methods, achieving 93.7% of mAP on CUHK-SYSU dataset and 46.4% of mAP on PRW dataset, respectively.

Keywords: Person search, Person occlusion, Attention mechanism, Anchor-free

1 Introduction

Person search refers to the simultaneous location and identification of the pedestrian to be queried from surveillance videos given a particular pedestrian picture, which is also known as the unified task of pedestrian detection and person re-identification [1, 2, 3]. It requires not only addressing occlusions, pose changes, and background clutter, but also performing detection and re-identification simultaneously by using a unified and optimized framework. The results suggest that independent optimization may delete helpful contextual information for the re-identification network, while joint optimization can achieve better results [4].

In recent years, with the gradual implementation of various social security projects such as Internet+ and smart city, the potential application value of camera networks has received extensive attention [5, 6]. At the same time, with the increase in the flow of people, social security problems are becoming more and more serious. Traditional feature recognition technology cannot show good effect for fuzzy images and videos. Person

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

search methods can solve such problems well, and it has a vital role to play in the fields of crime retrieval, cross-camera personnel tracking, and personnel activity analysis.

Person search frameworks are mainly divided into three types: (1) two-step frameworks, which consider person re-identification and pedestrian detection as two independent tasks (as shown in Fig. 1a); (2) one-step two-stage frameworks, which utilize ROI Align of two-stage detectors to achieve end-to-end training for detection and re-identification (as shown in Fig. 1b); and (3) one-step one-stage frameworks, which focus on implementing the re-id task based on the anchor-free model. The current one-step one-stage framework achieves the best results in the person search domain, and its representative research results based on the anchor-free model are presented in the studies [7, 8]. In this paper, the jointly optimized one-step one-stage framework is chosen to design a person search network model (as shown in Fig. 1c).

The first anchor-free-based person search model is proposed to solve the misalignment problem at three different levels (scale, region, and task) through a feature alignment aggregation module [7]. The study shows that the channels of convolutional features have different degrees of activation response for different parts of the pedestrian [9]. The form of target occlusion can be described by different feature channels, so that different occlusion problems can be handled effectively. An efficient attention module CBAM is proposed [10], which is composed by a channel attention module and a spatial attention module. Visualization analysis reveals that the network model embedded in the CBAM module will pay attention more accurately on the correct object to be classified during the inference process, achieving better detection results.

Inspired by the above results, an anchor-free model (ABOS) based on the spatial attention module and the channel attention module is proposed in this paper. The TOIM

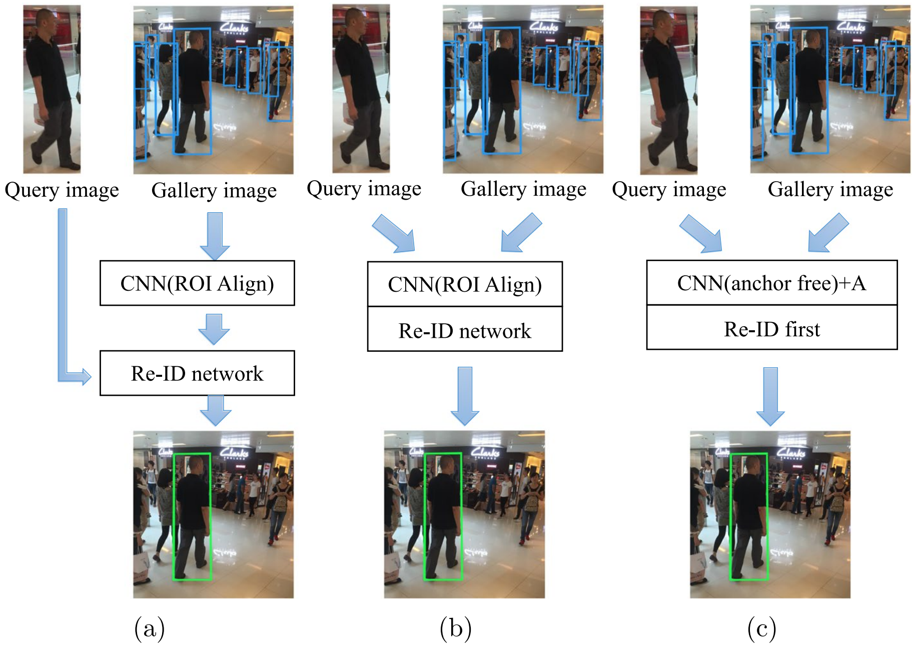

Fig. 1 Three different person search methods. **a** The traditional method first detects person in the gallery image and then estimates the re-identification features to find the target. **b** The end-to-end training method for joint re-identification and detection. **c** Our proposed ABOS model

loss optimizer that combines the advantages of triple loss function and online instance matching (OIM) loss function is chosen. Our method emphasizes the importance of difficult samples and simplifies the batch construction process of the triple loss function, greatly speeding up the convergence rate and effectively improving the precision of person re-identification. In order to validate the accuracy of the ABOS model, experiments are performed on the complete image datasets CUHK-SYSU and PRW with identity information and bounding box, respectively. The main contributions made by this paper are as follows.

- An anchor-free model (ABOS) that fuses the spatial attention module and the channel attention module is used to express different forms of occlusion and highlight the target regions of occluded pedestrians.
- By fusing the deep and shallow features of the attention modules, the network is guided to focus on the visible regions of the occluded targets. And the semantic information of pedestrians is extracted, thus improving the accuracy of person search when the occluded person is queried.
- Helping the proposed model learn better features and selecting information in the scene adaptively by introducing attention mechanism. Extensive experimental results show that the method proposed in this paper is not only fast and efficient, but also solves the problem of low accuracy of the one-stage network compared with the existing methods. Achieved mean average accuracy (mAP) is 93.7% and 46.4% of the original evaluation protocols provided in CUHK-SYSU and PRW datasets, respectively.

The rest of this paper is presented below. The related work on person search, pedestrian detection, and person re-identification is presented in Sect. 2. An attention-based one-stage framework for person search (ABOS) method is discussed in detail in Sect. 3. Experimental results are analyzed in Sect. 4. Finally, the work of the full paper is summarized and the future planning is discussed in Sect. 5.

2 Related work

This paper summarizes the previous research experience in person search, pedestrian detection, and person re-identification. An attention-based person search method (ABOS) is constructed by using optimal technology.

2.1 Person search

The person search problem was first introduced by Xu et al. [11], and a sliding window strategy is proposed to combine person re-identification and pedestrian detection to model common and unique characteristics of pedestrians, but the efficiency is low. Since the opening of the PRW and CUHK-SYSU datasets, the field of person search has attracted extensive attention and some methods have been proposed to improve its effectiveness. Zheng et al. [12] have discovered a cascade of fine-tuning strategies and confidence-weighted similarity metric to implement a two-step person search framework. Xiao et al. [1] proposed a combined OIM loss function and single neural network Faster R-CNN for joint training, which enhanced the feature recognition ability and realized a one-step two-stage end-to-end person search framework. Many improvement algorithms of this method were derived, such as multi-loss algorithm [13], which fuses pre-training and multi-loss; IEL algorithm [14], which refines the loss and improves the learned features for inaccurate pedestrian recognition robustness; and IAN algorithm [15], which uses a deeper network and adds a center loss (CL) function.

Some researchers approach pedestrian search from a new perspective. Liu et al. [16] repositioned the person search task as a no-detection process and used convolutional long short memory networks to recursively correct the position of pedestrian frames and match to obtain accurate pedestrian locating frames. Chang et al. [17] proposed the relational context-aware agents (RCAA) algorithm that the person search framework was introduced to the deep reinforcement learning for the first time. A cross-level semantic alignment (CLSA) approach for multi-scale pedestrian matching was proposed by Lan et al. [18]. The first anchor-free model improving the efficiency and simplicity of the one-step model with optimal results was introduced by Yan et al. [7].

We improve its network structure and increase the precision of person search based on the anchor-free model in this paper.

2.2 Pedestrian detection

Early pedestrian detection methods are mainly based on linear classifiers, handcraft features, integral channel features (ICF) [19], aggregated channel features (ACF) [20], and deformable part model (DPM) [21], etc. With the rapid development of deep learning methods in the field of object detection, methods based on convolutional neural networks can extract more discriminative features. Ouyang et al. [22] proposed joint deep model to deal with occlusion problem by jointly learning the visibility and features of different body parts. Tian et al. [23] proposed StrongParts model to deal with the occlusion problem. Zhang et al. [24] applied the general-purpose object detection method Faster R-CNN to the task of the pedestrian detection, but its operation speed was not satisfactory. The methods in [25, 26] achieved better results through various adjustments. Cai et al. [27] introduced a CompACT algorithm to learn CNN detector cascades.

The previously introduced methods are mainly two-stage detectors. In terms of one-stage detectors, a RetinaNet model was proposed by Lin et al. [28] with focal loss to address the class imbalance problem. The fused DNN detector proposed by Du et al. [29] used SSD detectors as pedestrian region candidate networks. Parallel multiple networks are used to optimize the candidate regions, and the semantic segmentation information is also incorporated into the detection process, effectively solving the small-scale and occlusion problems. Yan et al. [7] developed a one-stage-based anchor-free detector, achieving the best results.

Furthermore, the attention mechanism can also be introduced to solve the occlusion problem. Zhang et al. [30] proposed to express different forms of occlusion using the inter-channel attention mechanism, each of which can be expressed as a weighted combination between different channels.

In this paper, we adopt an anchor-free detector that combines with an attention mechanism to enhance feature extraction, to train the network to focus on the iconic part of the pedestrian, and to suppress the accuracy degradation problem caused by occlusion.

2.3 Person re-identification

Person re-identification aims to query pedestrians from a set of pedestrian candidates. Early person re-identification methods are mainly based on metric learning and artificially designed features.

Currently, CNN-based re-identification methods fall into two main categories: classification models and siamese models. These two types of models are often trained as feature extractors with siamese loss [31, 32], triplet loss [33], and cross-entropy loss [34, 35]. Cheng et al. [36] trained CNN models for maximizing the feature distances between different pedestrians and minimizing the feature distance between the same pedestrians by using triplet samples. Xiao et al. [37] classify identities to learn features while using a dual or triple loss function. The classification model proposed by Zheng et al. [12] achieves higher accuracy than the siamese model. Han et al. [38] proposed a novel proxy triplet loss, which solves the problem of inability to construct a standard triplet loss function in person search.

In recent years, attention [39, 40] mechanisms have been employed to learn better pedestrian recognition features. The PDC model [41] enhances pedestrian features with pose normalized images and re-weighting using channel attention. The HydraPlus-Net model [42] aggregates multiple feature layers within a spatial attention region. Xu et al. [31] solved the problems of person re-identification and pose estimation in a joint framework, where pose estimation results generate spatial attention maps and visibility scores.

As for the problem that re-identification is prone to regional dislocation in the anchor-free framework, this paper introduces both the channel attention mechanism and the spatial attention mechanism into the anchor-free framework and follows the "re-id first" principle to generate more discriminative and robust feature embedding and thus makes person search much more accurate.

3 Methods

This paper designs an attention-based one-stage framework (ABOS) for person search, which integrates a new anchor-free pedestrian detector and TOIM loss re-identification technology. The steps are as follows: (1) The anchor-free pedestrian detector in [43] is improved to integrate the attention module, which can fully guarantee the performance of the detection task; (2) TOIM loss [7] is integrated with the improved anchor-free pedestrian detector in this paper and follows the "re-id first" principle to effectively solve the problem of re-identification task in person search without generating additional re-id features; and (3) training and testing are performed on the CUHK-SYSU and PRW datasets, and the experimental results demonstrate the feasibility and advancement of the ABOS model.

3.1 Model structure

The ABOS framework is improved and designed based on the FCOS network [43]. As shown in Fig. 2, for the input image $I \in R^{3\times H\times W}$, we extract a set of features $\{C_1, C_2, C_3, C_4\}$ from ResNet-50. According to [44], the low-level features in the shallow convolutional layer retain the spatial information used to construct the edges of

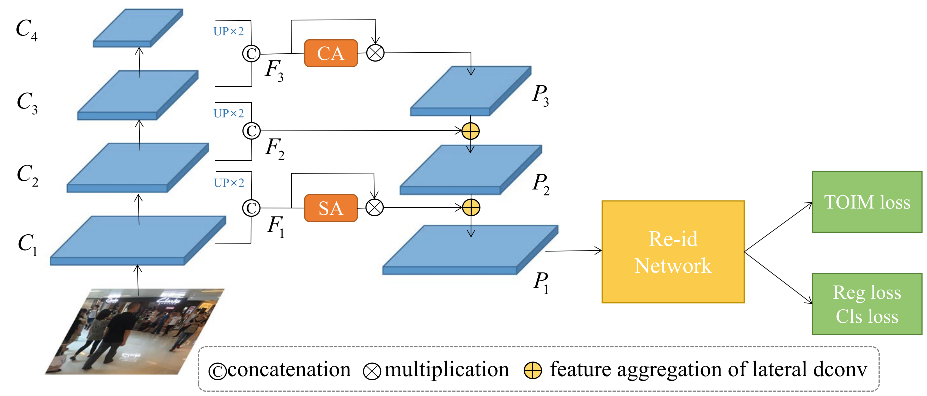

Fig. 2 Person search network structure

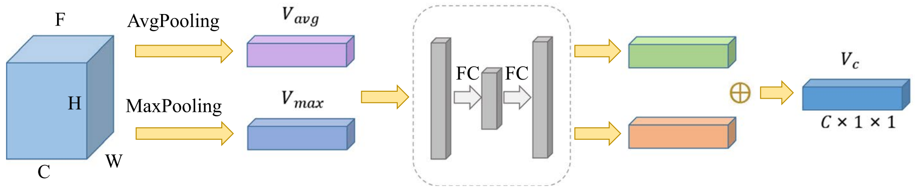

Fig. 3 Channel attention module

the object, while the deep convolutional layer retains the semantic information used to locate the target. Meanwhile, according to the inherent properties of the neural network, the extracted features are stratified into $F_1\{C_1, C_2\}$, $F_2\{C_2, C_3\}$, and $F_3\{C_3, C_4\}$. They are combined through operations such as concatenation and upsampling to preserve more information from different layers and then fused with the attention mechanisms and fed into the feature pyramid. To ensure that the scale of person search is aligned, we use 3×3 lateral deformable convolutions instead of 1×1 lateral convolutions for feature fusion. The first layer of the feature pyramid aggregates features of the multi-level feature maps in the backbone. For re-id, we only need to learn from the maximum output feature map P_1.

3.2 Attention module

In this paper, channel attention module and spatial attention module are used to enhance the more discriminative feature representation, respectively. For example, embed the feature layer $F_1\{C_1, C_2\}$ fused by upsampling and concatenation operations into a spatial attention module, while $F_3\{C_3, C_4\}$ is embedded into the channel attention module.

3.2.1 Channel attention module (CA)

The channel attention module is mainly concerned with "what" is meaningful in a given input image F, and each channel of the feature map is treated as a feature detector, generating a descriptive vector $V \in R^{C\times1\times1}$ that expresses the importance of the channel. As shown in Fig. 3, for the input feature map F_3, both average pooling and max pooling are used to aggregate the global information of each channel feature, which generates the average pooling feature description vector $V_{\text{avg}} \in R^{C\times1\times1}$ and the maximum pooling feature description vector $V_{\max} \in R^{C \times 1 \times 1}$, respectively. Then, the two description vectors are embedded in a shared network consisting of a hidden layer and a multilayer perceptron (MLP) to generate the channel attention vector V_c. The calculation formula is as follows:

$$\begin{aligned} V_c &= \sigma\left(MLP\left(AvgPool(F_3)\right) + MLP(MaxPool(F_3))\right) \\ &= \sigma\left(W_1\left(\delta\left(W_0 V_{\mathrm{avg}}\right)\right) + W_1(\delta(W_0 V_{\max}))\right) \end{aligned} \tag{1}$$

where δ denotes the ReLu activation function, σ denotes the sigmoid function, and $W_0 \in R^{C/r \times C}$ and $W_1 \in R^{C \times C/r}$ are the weights of the two fully connected layers of the MLP, where r denotes the dimensionality reduction ratio. F_{chn} is obtained by weighting F_3. The calculation formula is as follows:

$$F_{chn} = V_c \otimes F_3 \tag{2}$$

where $\otimes$ denotes the channel-by-channel multiplication.

3.2.2 Spatial attention module (SA)

The spatial attention module is mainly concerned with "where" of the information part, and spatial attention maps are generated from the spatial relationships between features. The spatial attention map is used to reactivate the input features so that the model focuses on obscuring the target pedestrians and suppressing the interference of the background in this paper. As shown in Fig. 4, for the input feature map F_1, first concatenate average pooling and maximum pooling to generate spatial description feature map $F_{\mathrm{avg}} \in R^{1 \times H \times W}$ and $F_{\max} \in R^{1 \times H \times W}$, respectively. And then generate the spatial attention map $M_s \in R^{1 \times H \times W}$ by using a 3×3 convolutional layer. The calculation formula is as follows:

$$\begin{aligned} M_s &= \sigma\left(f^{3 \times 3}\left([AvgPool(F_1); MaxPool(F_1)]\right)\right) \\ &= \sigma\left(f^{3 \times 3}\left(\left[F_{\mathrm{avg}}; F_{\max}\right]\right)\right) \end{aligned} \tag{3}$$

where $f^{3 \times 3}$ denotes the 3×3 convolutional layer and σ denotes the sigmoid function. The final feature map F_{sp} is obtained by reactivating the input feature map F_1, which is calculated as follows:

$$F_{sp} = M_s \odot F_1 \tag{4}$$

where $\odot$ denotes the element-by-element multiplication of the feature map.

3.3 Optimizer

In the study of person search, due to the regression inaccuracy and missed detection caused by the occlusion problem, it is necessary to narrow the feature differences of

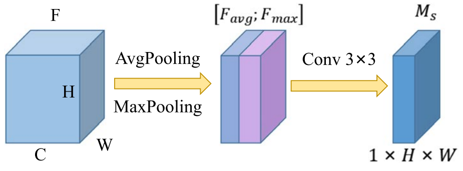

Fig. 4 Spatial attention module

the same pedestrian instance and expand the feature differences of different pedestrian instances. Since the person search dataset has large amount of training IDs with limited number of samples per ID, the triplet-aided online instance matching (TOIM) loss function is used in re-id, which combines the OIM loss function and the triplet loss function.

A lookup table (LUT) $V \in R^{D\times L}$ is defined in OIM to store all features of tagged identities, where L denotes the table size and D denotes the feature size. Also, define a circular queue $U \in R^{D\times Q}$ to store the unlabeled identity features, where Q denotes the size of the queue. Given a label i and input feature x, the probability that x is considered as the i-th feature with labeled identity according to the above two data structures is:

$$p_i = \frac{\exp\left(v_i^T x\right)/\tau}{\sum_{j=1}^{L} \exp\left(v_j^T x\right)/\tau + \sum_{k=1}^{Q} exp\left(u_k^T x\right)/\tau} \tag{5}$$

where $v_i^T x$ denotes the cosine similarity between x and the i-th labeled identity, $u_k^T x$ denotes the cosine similarity between x and the k-th unlabeled identity, and τ denotes the hyperparameter that controls the softness of the probability distribution. The goal of OIM is to maximize the expected log-likelihood function of feature x as:

$$L_{\text{OIM}} = E_x[\log p_t] \tag{6}$$

where t denotes the label of the labeled identity, and its gradient with respect to x can be derived as:

$$\frac{\partial L}{\partial x} = \frac{1}{\tau}\left[(1-p_t)v_t - \sum_{\substack{j=1\\ j\neq t}}^{L} p_j v_j - \sum_{k=1}^{Q} q_k u_k\right] \tag{7}$$

The set of candidate features with identity labels n and m is defined in triplet $X_n = \left\{x_{n,1}, \ldots, x_{n,S}, v_n\right\}$ and $X_m = \left\{x_{m,1}, \ldots, x_{m,S}, v_m\right\}$, where S denotes the number of features sampled for a pedestrian, v_i denotes the i-th feature in the LUT and $x_{i,j}$ denotes the j-th feature of the i-th pedestrian. The triplet loss function is calculated as follows:

$$L_{\text{tri}} = \sum_{\text{pos,neg}} \left[M + D_{\text{pos}} - D_{\text{neg}}\right] \tag{8}$$

where M denotes the boundary distance between negative and positive samples, D_{neg} denotes the distance between negative sample pairs, and D_{pos} denotes the distance between positive sample pairs. In summary, the TOIM loss function is calculated as follows:

$$L_{\text{TOIM}} = L_{\text{OIM}} + L_{\text{tri}} \tag{9}$$

4 Experiments and results analysis

This section describes the details of the experiment and the experimental environment.

4.1 Datasets

CUHK-SYSU [1] is a scene-diverse and large-scale person search dataset, containing 18184 images, 8432 identities, and 96143 bounding boxes with annotations with annotations (only 96131 after deduplication). The dataset is mainly derived from street photography and movies. The dataset is divided into training set and test set, the training set includes 11206 images, 5532 identities, and 55272 bounding boxes with annotations (only 55260 after deduplication), and a test set contains 6978 images, 2900 identities, and 40871 bounding boxes with annotations. A set of protocols ranging in size from 50 to 4000 are defined in the test set.

PRW [12] is a large-scale person search dataset, containing 11816 images, 932 identities, and 34304 bounding boxes with annotations. The dataset is acquired by six different cameras on a university campus. The dataset is divided into test set and training set, a test set contains 6112 images and 450 identities, and the training set contains 5704 images and 482 identities.

4.2 Evaluation metrics

This paper uses the mean average precision (mAP) and top-1 accuracy of the original evaluation protocol provided by the dataset to evaluate the people search performance of the ABOS model. Average precision (AP) summarizes the precision–recall curve for a category, which is the weighted average of the precision obtained at each confidence threshold, and the increase in recall is used as a weight compared with the previous confidence threshold, calculated as follows:

$$\mathrm{AP} = \sum_{n} (R_n - R_{n-1})P_n \tag{10}$$

where R_n and P_n represent the recall and precision of the n-th confidence threshold, respectively. mAP represents the average AP over all categories, and top-1 represents the accuracy of the category with the first predicted probability matching the actual result.

4.3 Experimental details

In this paper, PyTorch and MMDetection tools are used to implement an attention-based one-stage framework for person search (ABOS). The backbone network uses ResNet-50 pre-trained on ImageNet, taking 4 layers $\{C_1, C_2, C_3, C_4\}$ from ResNet-50. The attention network consists of $\{C_1, C_2\} + \mathrm{SA}$, $\{C_2, C_3\}$, and $\{C_3, C_4\} + \mathrm{CA}$, and detailed information is shown in Table 1. The learning rate on NVIDIA Corporation GP104GL 8GB is set to 0.0001. τ in the OIM loss function is set to 0.1. Stochastic gradient descent (SGD) is used to optimize the batch size which is 1 of the network, and the weight decay is 0.0001. The training set includes the ground-truth bounding box and the detection bounding box

Table 1 Layered fusion implementation details

Attention network layer	n×c×h×w,attention
P_1	1×768×h×w,A,M,3×3C,Sigmoid
P_2	1×1536×h×w
P_3	1×3072×h×w,A,M,1×1F,ReLu,1×1F,Sigmoid

n denotes the number of filters, c denotes the channels of the filter, h/w denotes the size of the filter, A denotes the average pooling layer, M denotes the maximum pooling layer, C denotes the convolutional layer, and F denotes the fully connected layer of the training image. During the training process, the long edges of the images are randomly adjusted from 667 to 2000, and the images in the test set images are adjusted to 1500×900 during the testing process.

4.4 Comparison with other attention models

The effectiveness of fused attention networks in person search applications in recent years is compared in Table 2. IDE+ATT-part [9] proposed a self-matching speed learning method based on attention guided to balance different occlusion levels. DHFF [45] treats the shallow network as an attention network, which achieved mAP 90.2% and top-1 91.7% on the CUHK-SYSU dataset, and mAP 41.1% and top-1 70.1% on the PRW dataset. QEEPS [4] proposed an improved QSSE-Net based on Squeeze-and-Excitation attention network, which achieved mAP 88.9% and top-1 89.1% on the CUHK-SYSU dataset, and mAP 37.1% and top-1 76.7% on the PRW dataset. Obviously, ABOS model fused with attention networks achieves the best effect.

4.5 Comparison with prior art

The anchor-free model is used as a baseline for comparison experiments on the CUHK-SYSU and PRW datasets to verify the effectiveness of the attention network designed in the paper.

The workstation graphics card configuration is NVIDIA Corporation GP104GL 8GB. Due to the limited hardware conditions in the laboratory, the parameter configuration is reduced to reproduce the project [7], reducing the learning rate and weight by 10 times, and reducing the batch size to 1.

Comparison of results on the CUHK-SYSU dataset: The evaluation metrics mAP and top-1 obtained by reducing the parameter configuration to reproduce the project [7] are 92.5% and 93.0%. Respectively, on this basis, the proposed network model is implemented. As shown in Table 3, it is clear that the ABOS model is improved on the CUHK-SYSU dataset, where the mAP is 93.7% and top-1 is 94.3%. Comparing with the first anchor-free person search model AlignPS [7], our proposed model has a relative improvement of 1.2% on mAP and 1.3% on top-1.

Comparison of results on the PRW dataset: The evaluation metrics mAP and top-1 obtained by reducing the parameter configuration to reproduce the project [7] are 45.4% and 83.3%, respectively. As shown in Table 3, the proposed model improves the mAP metric and top-1 metric for person search on PRW dataset, where mAP is 46.4% and

Table 2 Comparison of attention network models

Models	CUHK-SYSU		PRW	
	mAP(%)	**Top-1**(%)	**mAP**(%)	**top-1**(%)
IDE+ATT-part [9]	83.8	84.0	36.5	79.1
QEEPS [4]	88.9	89.1	37.1	76.7
DHFF [45]	90.2	91.7	41.1	70.1
Ours	93.7	94.3	46.4	84.9

Table 3 Comparison of network models implemented on CUHK-SYSU and PRW datasets for person search

	Models	**CUHK-SYSU**		**PRW**	
		mAP (%)	**Top-1** (%)	**mAP** (%)	**Top-1** (%)
One-step	OIM [1]	75.5	78.7	21.3	49.4
	RCAA [17]	79.3	81.3	–	–
	BINet [46]	90.0	90.7	45.3	81.7
	NAE [47]	91.5	92.4	43.3	80.9
	NAE+ [47]	92.1	92.9	44.0	81.1
	AlignPS [7]	92.5	93.0	45.4	83.3
	AlignPS+ [7]	92.9	93.6	45.9	83.8
	Ours	93.7	94.3	46.4	84.9
	Ours+	93.4	94.0	45.2	84.7
Two-step	CNN+CLSA [18]	87.2	88.5	38.7	65.0
	OR [48]	92.3	93.8	52.3	71.5
	FPN+RDLR [38]	93.0	94.2	42.9	70.2

AlignPS [7] and AlignPS+ [7] represent the effect achieved when reducing the parameter to replication

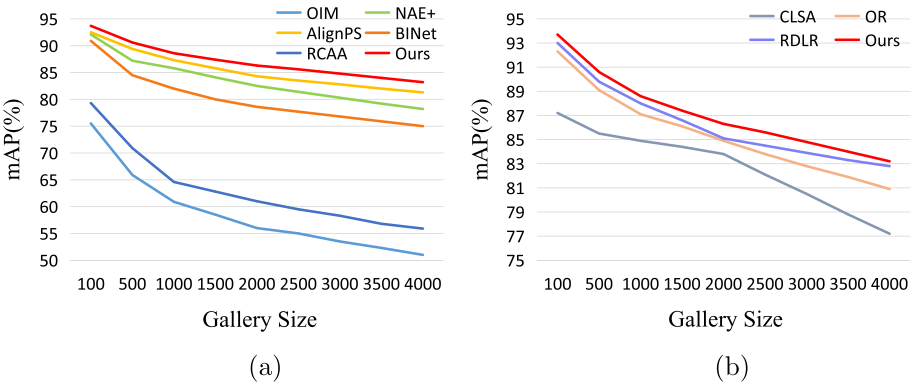

Fig. 5 Effect of different gallery sizes on mAP on the CUHK-SYSU dataset. **a** The model comparison graph of the one-step. **b** The model comparison graph of the two-step

top-1 is 84.9%. Comparing with the model AlignPS, our proposed model has a relative improvement of 1% on mAP and 1.6% on top-1.

A comparison of the results of different methods is shown in detail in Fig. 5. Person search is more challenging as the gallery size increases, and the person search has the best performance when the gallery size is equal to 100. It can be concluded that our method is superior most person search methods compared to state-of-the-art person search techniques today. Therefore, the proposed model can be used in practical applications.

5 Results and discussion

To address the pedestrian occlusion problem commonly faced in person search, this paper proposes an attention-based one-stage framework for person search (ABOS). Introducing spatial attention mechanism and channel attention mechanism into the multi-scale feature fusion, the channel attention module establishes the correlation between channels in the deep feature map, while the spatial attention module extracts the spatial information of the shallow feature map to highlight the obscured pedestrian targets. The TOIM loss is chosen in the re-identification module to further enhance the learning ability of the network for pedestrian similarity and the distinguishability of pedestrians. In this paper, the two tasks of person re-identification and pedestrian detection are fused in a network model for joint modeling optimization, thus improving the detection accuracy and recognition rate when the person to be queried is obscured. To verify the effectiveness of the proposed method, training and testing are carried out on CUHK-SYSU and PRW datasets. Numerous experiments have shown that the ABOS model outperforms most existing person search models and achieves high accuracy.

Due to the complexity of surveillance scenarios, person search is still a long way from practical applications. In the future, we will be working on the following four areas.

- Due to the limited hardware conditions in the laboratory, the proposed model fails to achieve the best results. We will apply a graphics card with a higher configuration and find the optimal parameters through continuous experiments to achieve the best results of the model in the future.
- Although the person search model proposed in this paper achieves good results for person search in general scenarios, misjudgment still occurs in some special scenarios. For example, the search results do not match the target pedestrians in rainy or foggy days with low visibility. In the future, we will further investigate the person search methods in different environments [49].
- Our proposed method is only for images. How to search for target pedestrians in a video requires further refinement of the model and testing in the future.
- The person search methods nowadays generally have not been widely used in the field of security, etc. and we still need to carry out a lot of theoretical research and equipment debugging [50–52].

Abbreviation

ABOS An attention-based one-stage framework for person search

Acknowledgements

All authors appreciate the efforts of the reviewers and editors who provided professional comments and suggestions.

Author contributions

All authors read and approved the final manuscript.

Funding

This research is supported by the National Natural Science Foundation of China under Grant 61873160, Grant 61672338, Natural Science Foundation of Shanghai under Grant 21ZR1426500, the Top Innovative Talents Program for Graduate students under Grant 2021YBR008 and National key research and development program of China under Grant2021YFC2801000.

Availability of data and materials

Corresponding authors can provide datasets analyzed during the current study on rational demand.

Declarations

Competing interests

There are no competing financial interests.

Received: 4 June 2022 Accepted: 15 August 2022
Published online: 03 September 2022

References

1. T. Xiao, S Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3376–3385 (2017)
2. S. Zhang, R. Benenson, B. Schiele, Citypersons: A diverse dataset for pedestrian detection, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4457–4465 (2017)
3. K. Zheng, W. Liu, L. He, T. Mei, J. Luo, Z.-J. Zha, Group-aware label transfer for domain adaptive person re-identification, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5306–5315 (2021)
4. B. Munjal, S. Amin, F. Tombari, F. Galasso, Query-guided end-to-end person search, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 811–820 (2019)
5. D. Deif, Y. Gadallah, A comprehensive wireless sensor network reliability metric for critical internet of things applications. *EURASIP J. Wirel. Commun. Netw.* 145 (2017)
6. F. Zhao, X. Sun, H. Chen, R. Bie, Outage performance of relay-assisted primary and secondary transmissions in cognitive relay networks. *EURASIP J. Wirel. Commun. Netw.* **60** (2014)
7. Y. Yan, J. Li, J. Qin, S. Bai, S. Liao, L. Liu, F. Zhu, L. Shao, Anchor-free person search, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7686–7695 (2021)
8. W. Liu, S. Liao, W. Ren, W. Hu, Y. Yu, High-level semantic feature detection: a new perspective for pedestrian detection, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5182–5191 (2019)
9. S. Zhang, D. Chen, J. Yang, B. Schiele, Guided attention in CNNs for occluded pedestrian detection and re-identification. *Int. J. Computer Vis.* (4) (2021)
10. S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in *European Conference on Computer Vision*, pp. 3–19 (2018)
11. Y. Xu, B. Ma, R. Huang, L. Lin, Person search in a scene by jointly modeling peoplecommonness and person uniqueness, in *The ACM International Conference* (2014)
12. L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3346–3355 (2017)
13. H. Liu, W. Shi, W. Huang, Q. Guan, A discriminatively learned feature embedding based on multi-loss fusion for person search, in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1668–1672 (2018)
14. W. Shi, H. Liu, F. Meng, W. Huang, Instance enhancing loss: Deep identity-sensitive feature embedding for person search, in *IEEE International Conference on Image Processing*, pp. 4108–4112 (2018)
15. J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, J. Feng, IAN: the individual aggregation network for person search. Pattern Recogn. **87**, 332–340 (2019)
16. H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, S. Yan, Neural person search machines, in *IEEE International Conference on Computer Vision*, pp. 493–501 (2017)
17. X. Chang, P.-Y. Huang, Y.-D. Shen, X. Liang, Y. Yang, A.G. Hauptmann, RCAA: Relational context-aware agents for person search, in *European Conference on Computer Vision*, pp. 86–102 (2018)
18. X. Lan, X. Zhu, S. Gong, Person search by multi-scale matching, in *European Conference on Computer Vision*, pp. 553–569 (2018)
19. P. Dollár, Z. Tu, P. Perona, S. Belongie, Integral channel features, in *British Machine Vision Conference, BMVC 2009*, London, UK, September 7–10, 2009. Proceedings (2009)
20. P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection. IEEE Trans. Pattern Anal. Mach. Intell. **36**(8), 1532–1545 (2014)
21. P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
22. W. Ouyang, X. Wang, A discriminative deep model for pedestrian detection with occlusion handling, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3258–3265 (2012)
23. Y. Tian, P. Luo, X. Wang, X. Tang, Deep learning strong parts for pedestrian detection, in *IEEE International Conference on Computer Vision*, pp. 1904–1912 (2015)
24. L. Zhang, L. Lin, X. Liang, K. He, Is faster R-CNN doing well for pedestrian detection, in *European Conference on Computer Vision*, pp. 443–457 (2016)
25. J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast R-CNN for pedestrian detection. IEEE Trans. Multimedia **20**(4), 985–996 (2018)
26. S. Zhang, L. Wen, X. Bian, Z. Lei, S.Z. Li, Occlusion-aware R-CNN: detecting pedestrians in a crowd, in *European Conference on Computer Vision*, pp. 657–674 (2018)
27. Z. Cai, M. Saberian, N. Vasconcelos, Learning complexity-aware cascades for deep pedestrian detection. IEEE Trans. Pattern Anal. Mach. Intell. **42**(9), 2195–2211 (2020)
28. T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 318–327 (2020)
29. X. Du, M. El-Khamy, J. Lee, L. Davis, Fused DNN: a deep neural network fusion approach to fast and robust pedestrian detection, in *IEEE Winter Conference on Applications of Computer Vision*, pp. 953–961 (2017)
30. S. Zhang, J. Yang, B. Schiele, Occluded pedestrian detection through guided attention in CNNs, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6995–7003 (2018)
31. J. Xu, R. Zhao, F. Zhu, H. Wang, W. Ouyang, Attention-aware compositional network for person re-identification, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2119–2128 (2018)
32. H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, End-to-end comparative attention networks for person re-identification. IEEE Trans. Image Process. **26**(7), 3492–3506 (2017)
33. S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification. Pattern Recogn. **48**(10), 2993–3003 (2015)
34. X. Fan, W. Jiang, H. Luo, M. Fei, Spherereid:deep hypersphere manifold embedding for person re-identification. J. Vis. Commun. Image Represent. **60**, 51–58 (2019)
35. W. Xiang, J. Huang, X. Qi, X. Hua, L. Zhang, Homocentric hypersphere feature embedding for person re-identification, in *IEEE International Conference on Image Processing*, pp. 1237–1241 (2019)

36. D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1335–1344 (2016)
37. T. Xiao, H. Li, x W. Li, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1249–1258 (2016)
38. C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang, C. Gao, N. Sang, Re-id driven localization refinement for person search, in *IEEE/CVF International Conference on Computer Vision*, pp. 9813–9822 (2019)
39. Z. Guo, D. Han, Sparse co-attention visual question answering networks based on thresholds. Appl Intell. (2022). https://doi.org/10.1007/s10489-022-03559-4
40. C. Chen, D. Han, C.-C. Chang, CAAN: Context-Aware attention network for visual question answering. Pattern Recognit. (2022). https://doi.org/10.1016/j.patcog.2022.108980
41. C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, in *IEEE International Conference on Computer Vision*, pp. 3980–3989 (2017)
42. X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, X. Wang, Hydraplus-net:attentive deep features for pedestrian analysis, in *IEEE International Conference on Computer Vision*, pp. 350–359 (2017)
43. Z. Tian, C. Shen, H. Chen, T. He, FCOS: Fully convolutional one-stage object detection, in *IEEE/CVF International Conference on Computer Vision*, pp. 9626–9635 (2019)
44. T. Zhao, X. Wu, Pyramid feature attention network for saliency detection, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3080–3089 (2019)
45. Y. Lu, Z. Hong, B. Liu, W. Li, N. Yu, DHFF: Robust multi-scale person search by dynamic hierarchical feature fusion, in *IEEE International Conference on Image Processing*, pp. 3935–3939 (2019)
46. W. Dong, Z. Zhang, C. Song, T. Tan, Bi-directional interaction network for person search, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2836–2845 (2020)
47. D. Chen, S. Zhang, J. Yang, B. Schiele, Norm-aware embedding for efficient person search, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12612–12621 (2020)
48. H. Yao, C. Xu, Joint person objectness and repulsion for person search. IEEE Trans. Image Process. **30**, 685–696 (2021)
49. W. Liang, J. Long, K.-C. Li, J. Xu, N. Ma, X. Lei, A fast defogging image recognition algorithm based on bilateral hybrid filtering. ACM Trans. Multimedia Comput Commun Appl. **17**(42), 1–16 (2021)
50. M. Cui, D. Han, J. Wang, An Efficient and safe road condition monitoring authentication scheme based on fog computing. IEEE Internet Things J. **6**(5), 9076–9084 (2019)
51. M. Cui, D. Han, J. Wang, K.-C. Li, C.-C. Chang, ARFV: An Efficient Shared Data Auditing Scheme Supporting Revocation for Fog-Assisted Vehicular Ad-Hoc Networks. IEEE Trans Vehicular Technol. **69**(12), 15815–15827 (2020)
52. H. Li, D. Han, M. Tang, A privacy-preserving charging scheme for electric vehicles using blockchain and fog computing. IEEE Syst J. **15**(3), 3189–3200 (2020)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com