Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.042308

ARTICLE

DAAPS: A Deformable-Attention-Based Anchor-Free Person Search Model

Xiaoqi Xin*, Dezhi Han and Mingming Cui

School of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China

*Corresponding Author: Xiaoqi Xin. Email: xinxiaoqi@stu.shmtu.edu.cn

Received: 25 May 2023 Accepted: 14 August 2023 Published: 29 November 2023

ABSTRACT

Person Search is a task involving pedestrian detection and person re-identification, aiming to retrieve person images matching a given objective attribute from a large-scale image library. The Person Search models need to understand and capture the detailed features and context information of smaller objects in the image more accurately and comprehensively. The current popular Person Search models, whether end-to-end or two-step, are based on anchor boxes. However, due to the limitations of the anchor itself, the model inevitably has some disadvantages, such as unbalance of positive and negative samples and redundant calculation, which will affect the performance of models. To address the problem of fine-grained understanding of target pedestrians in complex scenes and small sizes, this paper proposes a Deformable-Attention-based Anchor-free Person Search model (DAAPS). Fully Convolutional One-Stage (FCOS), as a classic Anchor-free detector, is chosen as the model's infrastructure. The DAAPS model is the first to combine the Anchor-free Person Search model with Deformable Attention Mechanism, applied to guide the model adaptively adjust the perceptual. The Deformable Attention Mechanism is used to help the model focus on the critical information and effectively improve the poor accuracy caused by the absence of anchor boxes. The experiment proves the adaptability of the Attention mechanism to the Anchor-free model. Besides, with an improved ResNeXt+ network frame, the DAAPS model selects the Triplet-based Online Instance Matching (TOIM) Loss function to achieve a more precise end-to-end Person Search task. Simulation experiments demonstrate that the proposed model has higher accuracy and better robustness than most Person Search models, reaching 95.0% of mean Average Precision (mAP) and 95.6% of Top-1 on the CUHK-SYSU dataset, 48.6% of mAP and 84.7% of Top-1 on the Person Re-identification in the Wild (PRW) dataset, respectively.

KEYWORDS

Person Search; anchor-free; attention mechanism; person detection; pedestrian re-identification

1 Introduction

The Person Search aims to locate and detect all pedestrian targets in a given image or video, providing their positions and related information. The Person Search model usually includes two tasks person detection and person re-identification (re-ID) [1]. The primary purpose of pedestrian detection is automatically detecting and locating pedestrians in images or videos. And person re-ID refers to the task of matching different images of the same pedestrian to their corresponding identity embeddings through deep learning. The Person Search model is more complex and has higher practical value due to its involvement in detecting, recognizing, and inferring relationships among multiple individuals [2].

In practical application, the difficulty of improving the accuracy of the person-finding task focuses on the fine-grained understanding of the image. Distinguishing similar targets requires detailed analysis and comparison of their detailed features and minimizes interference from complex background factors such as changes in the appearance of pedestrian targets and crowds. Therefore, algorithms with strong robustness, high accuracy, and high efficiency need to be designed for Person Search tasks to cope with these challenges.

Currently, mainstream deep learning-based Person Search models typically utilize neural networks to learn image features and then perform detection through object classification and position regression. This type of model can be further divided by model characteristics into one-stage, two-stage, and one-step two-stage Person Search models [3–8], as shown in Fig. 1.

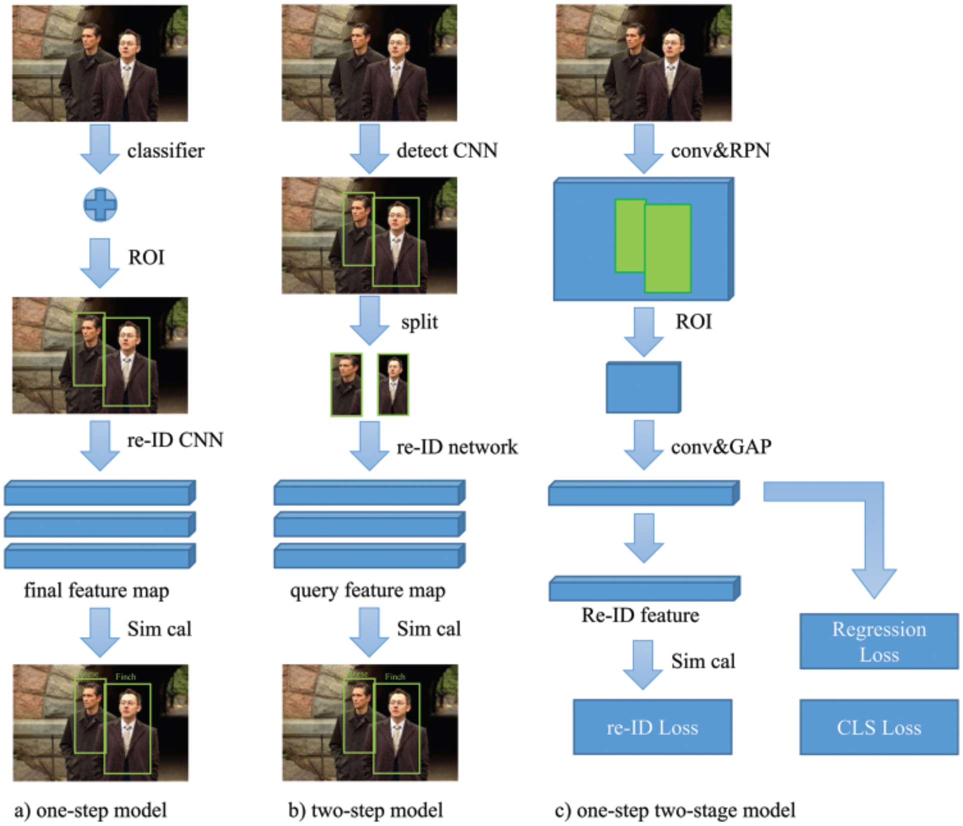

Figure 1: Classification and comparison of three person search models

(a) DAAPS: A Deformable-Attention-Based Anchor-Free Person Search Model. The one-step Person Search model [9], also known as end-to-end Person Search models, directly outputs pedestrian targets' position and size information from the input image. This model type usually has a faster speed and can achieve real-time detection. However, its detection accuracy is relatively lower due to the need for an explicit candidate region generation process. (b) Two-step Person Search model generates candidate boxes, then performs classification and position regression to obtain the final detection results. This model type usually has higher detection accuracy but needs to process many candidate regions, resulting in high computational complexity and relatively slow speed. (c) One-step two-stage model employs a Region of Interest Align (ROI-Align) layer to aggregate features from the detected bounding boxes, allowing detection and re-ID to share these features. Our model adopts a two-stage detector, such as Faster Region-based Convolutional Neural Networks (R-CNN) [10].

Anchors [10] are commonly used prior information in object detection tasks, which set a few fixed sizes and aspect ratios of anchor boxes to predict the position and size of the targets. The existence of anchor boxes can improve the accuracy of the Anchor-based model, but it is also affected by over-parameterization, which requires manual operation. Though, models based on Anchor-free [3,4] do not require predefined anchors but directly detect, extract, and recognize the human regions in the image. The model based on Anchor-free does not rely on prior boxes and can directly regress the position and size of the target, thus improving the computational efficiency.

The model based on Anchor-free has received much research due to its simple and fast construction. Models based on Anchor-free introduce an Aligned Feature Aggregation (AFA) module improved on the FCOS object detection framework. The FCOS architecture adopts the basic Residual Network (ResNet) backbone network structure and Feature Pyramid Network (FPN) to fuse multi-scale features and then deploys a decoupled detection head to detect targets on each scale separately. The network structure is shown in Fig. 2. AFA reshapes some modules of FPN by utilizing deformable convolutional kernels for feature fusion to generate more robust re-identification embeddings, overcoming the problem of models based on Anchor-free being unable to learn and detect aligned features for specific regions.

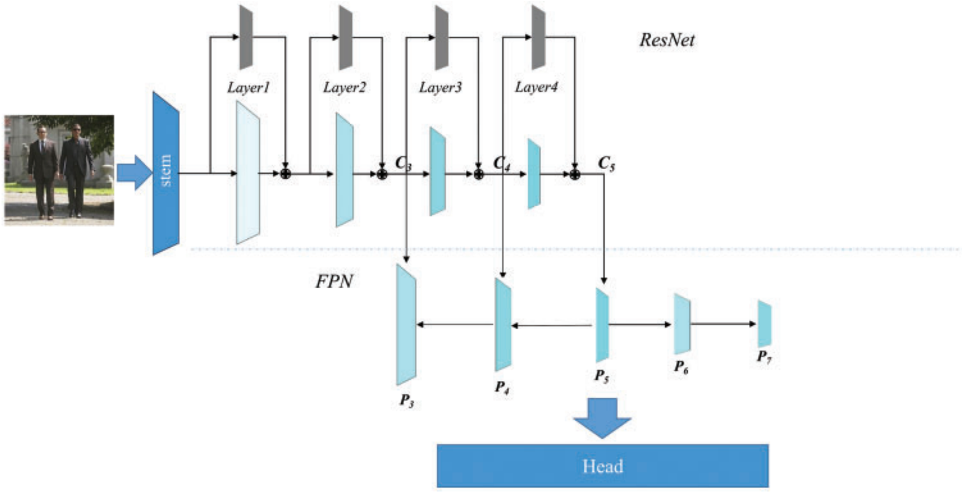

Figure 2: FCOS network architecture

The Attention Mechanism is one of the widely applied Computer Vision tasks to improve the utilization of crucial information by dynamically assigning importance or weight to different input parts in deep learning models. Common Attention Mechanisms include Spatial Attention, Channel Attention, and Self-Attention. The Deformable Attention [11] mechanism, as an extension of the Self-Attention mechanism, can learn feature deformations and adaptively adjust feature sampling positions to better handle target deformations and pose changes.

Besides, the design of loss functions is a crucial aspect in improving Person Search models. Common loss functions include Triplet Loss, Online Instance Matching (OIM) [1] Loss, etc. The

TOIM Loss function deployed in the DAAPS model combines the above two loss functions to match instances across frames or videos online accurately [12].

This paper proposes a Deformable-Attention-based Anchor-free Person Search (DAAPS) model inspired by the above research methods. Our main contributions are as follows:

This paper proposes a novel Person Search model based on an Anchor-free detector that incorporates the Deformable Attention mechanism for the first time. The proposed model first extracts image features by combining the Deformable Attention mechanism and convolutional layer, aligns the features by the AFA module, uses the FCOS detection head for target detection, and then feeds into the pedestrian re-recognition module to combine the character features with labels to get the search results.

The improved anchor-free detection feature extraction network, ResNeXt+, adds network branches and enhances the model scalability. The group convolution structure of ResNeXt+ can better extract multi-scale features, making the model more adaptable to complex Person Search tasks. Furthermore, the TOIM Loss function, a more suitable function, is chosen to better adapt to target variations, thus improving the model's detection accuracy.

To demonstrate that the optimization can help the model better understand the images at a finer granularity, the paper conducts extensive ablation experiments, in which mAP and top-1 are 95.0% and 95.6% on the CUHK-SYSU dataset, and 48.6% and 84.7% on the PRW dataset, respectively. The experimental results show that the DAAPS model outperforms the current best models based on Anchor-free, fully demonstrating rationality and effectiveness. In addition to this, the study conducted ablation experiments on various parts of the model and proved that the proposed modifications and optimizations are more suitable for the anchor-free model, thus illustrating the robustness and superiority of the present model.

The remainder of this paper is structured as follows. Section 2 reviews related work on the Person Search model based on Anchor-free and Attention Mechanisms. In Section 3, the implementation of the DAAPS model is depicted in detail. Section 4 analyzes and compares the experimental results, demonstrating the effectiveness of our proposed model. Finally, the paper is summarized, and future research is discussed in Section 5. Table 1 contains common abbreviations used in the paper for reference purposes.

Table 1: Table of common abbreviations

Abbreviation	Full name
FCOS	Fully convolutional one-stage
TOIM	Triplet-based online instance matching
PRW	Person re-identification in the wild
ResNet	Residual network
ResNeXt	The next dimension of ResNet
mAP	Mean average precision
re-ID	Re-identification
ROI-align	Region of interest align
R-CNN	Region-based convolutional neural networks
AFA	Aligned feature aggregation
FPN	Feature pyramid network

(Continued)

Table 1 (continued)

Abbreviation	Full name
AlignPS	The feature-aligned person search network
SGD	Stochastic gradient descent

2 Related Work

This section mainly reviews the existing Person Search models split by Anchor and based on Attention Mechanism, respectively, to highlight proposed model.

In this paper, databases such as Institute of Electrical and Electronics Engineers Xplore, Web of Science, the Engineering Index, ScienceDirect, Springer, and Arxiv are used as the main target for the Person Search model within the last five years through Google Scholar. We use various combinations of characters as search terms, e.g., “Person Search model”, “Anchor-free”, “pedestrian re-ID” and “Attention Mechanism”. After screening the suitable 54 papers whose sources include conferences, journals and online published papers were used as references for this paper. Deep learning models are gradually becoming one of the main targets of cyber attack. Attacks include adversarial attack, model spoofing attack, backdoor attack [13–15] and so on. How to reduce the impact of attacks and enhance robustness is also one of the focuses of model design.

2.1 Person Search Models Split by Anchor

The Person Search, object detection, and person recognition models have developed dramatically with the in-depth study of deep learning. Faster R-CNN is a classic two-step target detection model with Anchor, which can also be used for Person Search [10,16,17]. Chen et al. [18] combined Faster R-CNN and Mask R-CNN to search for people using two parallel streams. One stream is used for object detection and feature extraction, and the other is used to generate semantic segmentation masks for pedestrians to improve the accuracy of pedestrian searches further. He et al. [19,20] implemented a Siamese architecture instead of one stream for an end-to-end training strategy. The detection module is optimized based on Faster-RCNN. However, when the human object is occluded or deformed, the anchor point cannot accurately capture the shape and position information of the object, thus affecting the detection performance of the Anchor-based models.

Anchor-free detection is widely used in image detection [3,4,21–25], but it has been proposed to be applied to the Person Search model recently. Law et al. [26] proposed the earliest anchor-free target detection model, CornerNet, which does not rely on anchor boxes for target detection, but converts the target detection problem into a task of object corners detection. Subsequently, many classic Anchor-free detection models are proposed [3,4,9,21]. Yan et al. [4] proposed the AlignPS model based on FCOS that introduces Anchor-free into the Person Search task for the first time. In the AlignPS model, the AFA module addresses the issues of scale, area, and task misalignment caused by Anchor-free.

Nevertheless, as models based on Anchor-free usually rely on the prediction of critical points or center points, the precise positioning of targets is limited to some extent, and other methods are needed to help improve the model’s accuracy. There is still room for optimization in the accuracy of character detection and recognition and in the model architecture.

2.2 Person Search Models Based on Attention Mechanism

The application of the Attention Mechanism in the Person Search task can help improve the accuracy of detection and matching [21,24,27–34]. Chen et al. [21] introduced the channel attention module into the model based on Anchor-free to express different forms of occlusion and make full use of the spatial attention module to highlight the target area of the occlusion-covered objects. Zhong et al. propose an enhancement to feature extraction in their work by incorporating a position-channel dual attention mechanism [33]. This mechanism aims to improve the accuracy of feature representation by selectively attending to important spatial and channel-wise information. Zheng et al. introduce a novel hierarchical Gumbel attention network [34], which utilizes the Gumbel top-k re-parameterization algorithm. This network is designed for text-based person search and focuses on selecting semantically relevant image regions and words/phrases from images and texts. It enables precise matching by aligning and calculating similarities between these selected regions. Ji et al. develop a Description Strengthened Fusion-Attention Network (DSFA-Net) [35], which employs an end-to-end fusion-attention structure. DSFA-Net consists of a fusion and attention sub-network, leveraging three attention mechanisms. This architecture addresses the challenges in Person Search by enhancing the fusion of multimodal features and applying attention mechanisms to capture relevant information.

However, according to the experiments in this paper, Deformable Attention brings higher detection accuracy and is more suitable for Anchor-free Mechanism than channel attention and spatial attention. Cao et al. [36] and Chen et al. [37] proposed adding Deformable Attention Mechanism to Transformer [38] for the Person Search model. Although the Transformer works well for tasks such as long text or images, it has high computing and memory requirements due to the need to compute associations between all locations in the self-attention mechanism. Especially for longer input sequences, model training and reasoning can become more time-consuming and resource intensive. All above are why proposed model adopts the Deformable Attention mechanism to cooperate with the model based on an Anchor-free FCOS structure. Previous research has been limited to improving model performance by changing the detector or optimizing the re-identification algorithm. They focus only on the mechanisms they add and do not consider the effects of other attention mechanisms, nor do they validate the performance of the model under other attention mechanisms. The paper is the first to propose combining a deformable attention mechanism and an Anchor-free person search model with comparing with other attention, filling the gap in the impact of attention mechanism on the performance of pedestrian detection models. In addition, most of the previous studies have only considered the anchor frame and the attention mechanism itself, without considering how they are combined and what structures are needed to enable the two to be sufficiently combined to enhance model performance, which is one of the considerable differences between studies.

3 Method

In this section, the network architecture of DAAPS, the improved ResNeXt+ structure, the implementation of the deformable Attention Mechanism, and the calculation of the loss function are introduced in detail.

3.1 Network Architecture

The infrastructure of the proposed DAAPS model in this paper is designed based on FCOS. As shown in Fig. 3, for an input image of the size $I \in R^{3\times H\times W}$, the DAAPS model can simultaneously locate multiple target pedestrians in the image and learn re-ID embedding. Specifically, the proposed model first extracts image features and gets three levels of features according to the feature pyramid. A Deformable Attention Mechanism then processes it to handle better objects of different scales, directions, and shapes. Feature maps $\{P_3, P_4, P_5\}$ is obtained by down-sampling and weighting using strides of 8, 16, and 32. Subsequently, an AFA module is utilized to fuse features of different scales into a single embedding vector. The AFA module has multiple branches, each branch performing a weighted fusion of features at different scales and producing a fused and flattened feature vector. Then, an FCOS detection head is employed for object detection. It comprises two branches, namely the classification regression branch, each including four 3 × 3 deformable convolutional layers. The classification branch is utilized to classify each pixel's position determine if it is a queried object, and predict the object's category. At the same time the regression branch is employed to regress each pixel's position and predict the object's position and size.

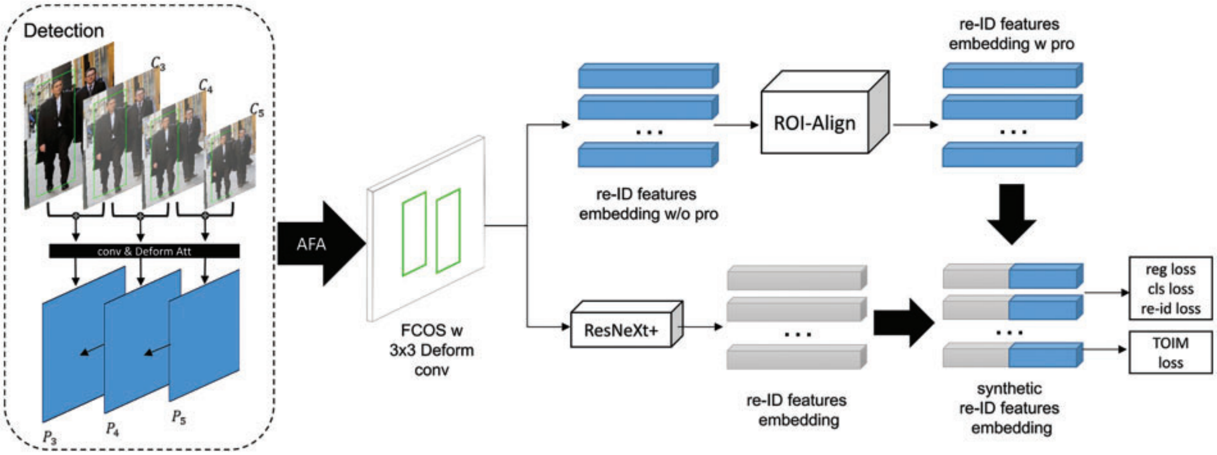

Figure 3: Network architecture of DAAPS

3.2 *ResNeXt+ Optimization*

The proposed DAAPS model is based on the classic model based on Anchor-free FCOS, which incorporates a single-stage object detection method and multi-scale feature fusion techniques. Unlike its predecessor, DAAPS introduces group convolution layers with more channels on top of the ResNet backbone to achieve ResNeXt [39] and deeper feature extraction. Subsequently, a pruning algorithm removes unimportant connections, reducing network complexity and improving the model's speed and accuracy. This improved network structure is referred to as the ResNeXt+ architecture in this paper.

For given input data of D dimensions $x = [x_1, x_2, \ldots, x_d]$, the corresponding filter weight is $w = [w_1, w_2, \ldots, w_d]$. A linearly activated neuron without bias can be expressed as:

$$\sum_{i=1}^{D} w_i x_i \tag{1}$$

That is, the data is split into individual features with low-dimensional embeddings. Each low-dimensional embedding undergoes a linear transformation and then aggregated using unit addition. It is a split-transform-merge structure that can be replaced with a more general function so that each structure makes use of the same topology. The aggregation transformation results are as follows:

$$\mathcal{F}(x) = \sum_{i=1}^{C} \mathcal{T}_i(x) \tag{2}$$

Among them, C is the size of the transformation set to be aggregated, namely cardinality. $\mathcal{T}_i$ (x) is any transformation, such as a series of convolution operations. ResNeXt+ is based on group convolutions, a strategy between regular convolutional kernels and depth-separable convolutions. By controlling the cardinality, a balance between the two strategies is achieved. The complete ResNeXt+ network structure is obtained, combined with the robust residual network. That is the addition of a shortcut is added to the simplified inception architecture, which is expressed as:

$$y = x + \sum_{i=1}^{C} \mathcal{T}_i(x) \tag{3}$$

ResNeXt+ adopts a Visual Geometry Group-like block stacking method, and the stacking process follows two principles: (a) If the same size spatial maps are produced, the blocks share the same hyperparameters. (b) Whenever the spatial map is downsampling twice, the width of the convolutional kernel is multiplied by 2. The structure is shown in Fig. 4.

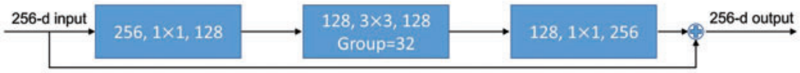

Figure 4: ResNeXt+ structure

That is, the input channel is reduced from 256 to 128 by a 1 × 1 convolutional layer and then processed using group convolution, with a convolutional kernel size of 3 × 3 groups of 32, and then up-dimensioned using a 1 × 1 convolutional layer. The output is added to the input to obtain the final output. ResNeXt+ employs group convolution to increase the width of the model. Compared to traditional convolution layers, group convolution splits the input features into several small groups, performs convolution operations on each small group separately, and finally concatenates all the convolution results of the small groups together.

Then pruning is implemented through the filter pruning algorithm [40] to optimize the network. Specifically, the L1 parametric was first used as the filter metric to determine which filters were more critical, and the L1 parametric was normalized to the average importance of each filter. A global pruning threshold is determined by setting the percentage of filter importance in the entire network, and the filters in the network are pruned according to the threshold. After pruning, the remaining filters are reattached to the network. Finally, fine-tuning is performed using a lower learning rate to recover the performance, and the fine-tuned network is initialized based on the weights before pruning. This approach allows the model to significantly reduce the number of parameters and the computational complexity of ResNeXt+ without losing much performance, resulting in a more efficient network.

ResNeXt+ has advantages over ResNet in terms of better performance and higher accuracy. Due to the ability of ResNeXt+ to simultaneously increase the depth and width of the model, it can better adapt to complex visual tasks. In addition, the more complex residual block is also employed in the ResNeXt+ structure, further improving the model's nonlinear expression ability. Therefore, the application of ResNeXt+ can improve the receptive field of FCOS and boost the effectiveness of the network. At the same time, ResNeXt+ has better generalization performance, making FCOS training and inference faster and more resource-efficient.

3.3 Deformable Attention Mechanism

For a given input feature mapping x ∈ $R^{C\times H\times W}$, where C is the number of channels, H and W are height and width, respectively. First of all, for each deformable convolution kernel m, the method computes the sampling offset Δp_{mqk} and attention weight distribution A_{mqk} for each sampled key k based on the linear projection of the query vector z_q and position vector p_q:

$$\Delta p_{mqk} = W^{(1)}_{mqk} z_q \tag{4}$$

$$A_{mqk} = softmax\left(W^{(2)}_{mqk} z_q\right) \tag{5}$$

where A_{mqk} is a scalar attention weight with a value range of [0, 1]. $W^{(1)}_{mqk}$ and $W^{(2)}_{mqk}$ are weights exploited to compute the offset and attention distribution in deformable convolutional kernels, which are both learnable parameters. Subsequently, the method multiplies the feature vector x_{mqk} located at $p_q + \Delta p_{mqk}$ in x by A_{mqk} to obtain a weighted feature vector y_{mqk}:

$$y_{mqk} = A_{mqk} \cdot x_{mqk} \tag{6}$$

$$x_{mqk} = W'_m x\left(p_q + \Delta p_{mqk}\right) \tag{7}$$

Finally, sum up y_{mqk} of all deformable convolution kernels to obtain the final output feature vector:

$$y = \sum_{m=1}^{M} \sum_{k=1}^{K} W_m y_{mqk} \tag{8}$$

To sum up, the calculation process of the Deformable Attention module includes three steps, calculation of sampling offsets and attention distributions, convolution on the input feature map, and weighted addition to obtaining the final output. The advantage of the Deformable Attention Mechanism lies in its ability to accurately capture long-range dependencies and geometric structures of the target, and exchange information between multi-scale feature maps, thus improving the accuracy of object detection and recognition.

3.4 Optimization Program

TOIM Loss function can be expressed as:

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

where L_{OIM} is OIM [1] Loss function. OIM aims to match the predicted instance in the image with the real instance. It first generates a pair of matching scores between the predicted and the real instance and then applies the matching score to calculate the loss function, which can be expressed as:

$$L_{OIM} = -\frac{1}{N} \sum_{i=1}^{N} log \frac{exp\left(w^T_{y_i} f_i / \tau\right)}{\sum_{j=1}^{K} I\left(y_j = y_i\right) exp\left(w^T_{y_i} f_i / \tau\right) + \sum_{j=1}^{K} I\left(y_j \neq y_i\right) exp\left(w^T_{y_j} f_i / \tau\right)} \tag{10}$$

Here, N is the batch size, K is the total number of classes, y_i represents the class label of sample i, f_i represents the feature vector of sample i, w_{y_i} represents the weight vector of class y_i, I represents the indicator function, and τ is the temperature parameter. OIM is designed to maximize the scores of the predicted instances of its underlying actual class and minimize the scores of other classes.

The L_{tri} stands for the Triplet Loss function, mainly deployed to address the association problem among multiple images of the same identity in pedestrian re-identification. L_{tri} works by encouraging the embeddings of positive instances to be closer while pushing the embeddings of negative instances away from the query target. Its computation is as follows. For a training sample, we need to select two images that are different from it, one belonging to the same category as the sample and the other belonging to a different category from the sample.

Given a training set containing m triples (A, P, N), where A represents the targeted sample, P represents the Positive sample, and N represents the Negative sample that does not belong to the same class as the targeted sample. The Triplet Loss function can be represented as follows:

$$L_{tri}(A, P, N) = \sum_{i=1}^{m} \max\left(\left\|f\left(A^{(i)}\right) - f\left(P^{(i)}\right)\right\|_2^2 - \left\|f\left(A^{(i)}\right) - f\left(N^{(i)}\right)\right\|_2^2 + \alpha, 0\right) \tag{11}$$

where $f(x)$ represents the embedding vector that maps the input sample x to the feature space. $A^{(i)}, P^{(i)}, N^{(i)}$ represent the anchor, positive, and negative samples in the *i* triplet, respectively. $\|\cdot\|_2$ is the norm of L_2. α is the margin, which indicates the minimum distance that should be different between positive and negative samples. By minimizing the Triplet Loss function, a face recognition model can be trained to map different photos of the same person to similar feature spaces.

4 Experiment

This section describes the experimental process and results from seven aspects, the datasets used in the experiment, the model's implementation details, the evaluation indicators, the attention ablation experiments, the comparison of experimental results, the loss function effect, the ResNeXt+ effect, and the visualization results.

4.1 Dataset

CUHK-SYSU [1] dataset is a large-scale pedestrian detection dataset with 1.14 gigabytes (GB) shared by the author Prof. Shuang Li. The images are sourced from two data types, authentic street snapshots and movies or TV shows. 12,490 images and 6,057 people to be detected are collected using hundreds of street scenes. Moreover, 5,694 images and 2,375 people to be detected are selected from movies or TV shows. Unlike the re-ID datasets that manually crop images of the queried person, the CUHK-SYSU is more closely related to real-world scenarios. The data is divided into training and testing sets, with the training set consisting of 11,206 images and 5,532 people to be queried and the testing set covering 6,978 images and 2,900 people to be queried. The images and people in the training and testing set are distinct.

PRW [41] dataset is an extension of the Market1501 dataset, typically employed for end-to-end pedestrian detection and person re-identification in raw video frames, and for evaluating person search and pedestrian re-ID in the wild. PRW dataset includes 11,816 video frames captured by six synchronized cameras and corresponding mat file annotations. Each mat file records the position of the bounding box within the frame along with its ID, and the dataset also contains 2,057 query boxes from the PRW dataset. The PRW dataset, available as an open-source repository on GitHub, is 2.67 GB. It encompasses a training set comprising 5,704 images along with 18,048 corresponding annotations and a test set containing 6,112 images with 250,062 annotations.

4.2 Implementation Details

The DAAPS model is implemented using PyTorch and the MMDetection toolkit. ResNeXt101+ serves as the backbone of our model. FPN with 3 × 3 deformable convolutions is applied as the neck, with a default Deformable Attention Mechanism. DAAPS is trained by the Stochastic Gradient Descent (SGD) optimizer, in which an initial learning rate is 0.0015, momentum is 0.9, and weight decay is set to 0.0005. Training and testing are conducted on an NVIDIA V100-SXM2-32 GB GPU. By default, the model is trained for 24 epochs. Linear learning rate warming up is chosen during training, with a warming up iteration of 1141 and a ratio of the warming up learning rate to the initial learning rate of 1/200. The learning rate is adjusted at the 16th and 22nd epochs; the remaining epochs are trained with the adjusted learning rate. During the training processes, the length of the image's longer side is adjusted to a random number between 667 and 2000, while the size of the images in the test set is adjusted to 1500 × 900 in the testing processes.

4.3 Evaluation Index

The performance of the Person Search model is evaluated through mAP and top-1 accuracy. mAP is one of the most commonly utilized evaluation metrics in object detection. First, for each category, all detection results are sorted by confidence from high to low. Then, based on the matching relationship between the actual and predicted categories of the detection results, the Precision-Recall curve is calculated at different confidence thresholds, and the area under the curve is obtained for each category. Finally, the AP values of all categories are averaged to obtain the mAP value of the model. The calculation formula is as follows:

$$mAP = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{R_c\left(m\right)}\sum_{i=1}^{R_C(m)}\left(\frac{i}{rank\left(i\right)}\sum_{j=1}^{i}TP\left(j\right)\right) \tag{12}$$

where C is the number of categories, R_c (m) is the number of class c in positive examples, rank (i) is the ranking of test results ranking i, and TP (j) represents the number of positive cases correctly detected among the first j detection results.

Top-1 refers to the top-1 accuracy of prediction results in classification tasks, namely, the highest probability value of correct results in the model prediction. ΔmAP stands for the change or delta in mean Average Precision. A positive ΔmAP indicates a performance improvement, while a negative ΔmAP suggests a decrease in performance. ΔTop-1 represents the change or delta in the Top-1 accuracy metric. It helps quantify the improvement or degradation in classification performance. The above four indexes are selected as indicators of the evaluation during experiments.

4.4 Ablation Studies of Attention Mechanism

To demonstrate the need for the Deformable Attention Mechanism, the ablation experiments are carried out on the CUHK-SYSU dataset, using a model based on Anchor-free without attention mechanisms as the Base. The impact of attention mechanisms on the algorithm is investigated, and the effectiveness of deformable attention mechanisms on DAAPS is verified. The experimental results are shown in Table 2.

Table 2: Effects of different attention mechanisms on the CUHK-SYSU dataset

Methods	mAP	Top-1	ΔmAP	ΔTop-1
Base [4]	93.1	93.4	–	–
CBAM [21]	93.7	94.3	+0.6	+0.9
DANet [42]	94.5	95.1	+1.4	+1.7
DAAPS	**95.0**	**95.6**	**+1.9**	**+2.2**

The addition of Convolutional Block Attention Module (CBAM) based on the Base, the combination of the Spatial Attention Mechanism and Channel Attention Mechanism, resulted in a slight increase in mAP and Top-1 index, 0.6% and 0.9%, respectively. Both mAP and Top-1 perform well if the combination is changed to the Attention Mechanism in Dual Attention network (DANet).

Nevertheless, none of these improvements compare to DAAPS, with the Deformable Attention Mechanism, giving mAP and Top-1 indexes a 1.9% and 2.2% boost on the CUHK-SYSU dataset. Our Deformable Attention Mechanism is more suitable for models based on Anchor-free.

To demonstrate the effectiveness and irreplaceability of the Deformable Attention Mechanism on the DAAPS model, this paper adds the CBAM Attention Mechanism to the model to work together with the Deformable Attention Mechanism. Theoretically, the superimposition of attention mechanisms can help the model learn the importance of different parts in the input and adjust the model's attention based on their importance to better capture relevant features. However, experimental data on the CUHK-SYSU dataset show that including CBAM results in a 0.4% decrease in both mAP and Top-1. Although the model's performance is only slightly reduced, it fully proves the robustness of the proposed model, as shown in Table 3.

Table 3: Prove the robustness of DAAPS on the CUHK-SYSU dataset

Methods	mAP	Top-1	ΔmAP	ΔTop-1
Base [4]	93.1	93.4	–	–
CBAM [21]	93.7	94.3	+0.6	+0.9
DAAPS w/o CBAM	**95.0**	**95.6**	**+1.9**	**+2.2**
DAAPS w/ CBAM	94.6	95.2	+1.5	+1.8

4.5 Comparison to State-of-the-Art

The results compared with state-of-the-art models are shown in Table 4, and our model outperforms most existing one-stage and two-stage Person Search models. The best result of the DAAPS model is compared to the previous best model, AlignPS+, where mAP has improved by 0.5% and Top-1 by 2.1% in the CUHK-SYSU dataset. This advantage is also reflected in the PRW dataset, where the mAP is 1.7% higher than the previous best task-consistent two-stage framework (TCTS) model. Moreover, our model is based on an Anchor-free network architecture, which runs faster than other models. Additionally, more efficient optimization algorithms and hyperparameter tuning techniques enable the proposed model to achieve better model performance with 24 training epochs.

Due to the limited training data in the PRW dataset and the fact that images are taken from six different camera viewpoints, the effect of all models on this dataset is constrained. Our model achieves the best mAP among all models. Although DAAPS's top-1 accuracy on the PRW dataset is 0.2% less than that of the current one-step best-performing ABOS, mAP is 2.1% higher. This indicates an improvement in the model's overall performance in terms of accuracy and recall, which are critical factors in tasks such as object detection or people search. The trade-off is reasonable because it leads to a more effective and comprehensive performance evaluation.

It can be seen that TCTS [43] achieves the highest accuracy in Top-1 on PRW, but it is a two-step model that requires a dedicated re-ID model to receive detection results for re-ID. As a model based on Anchor-free, DAAPS can adaptively learn the feature representation of the target without the need for a predefined fixed anchor point, which is certainly not affected by the selection and distribution of anchor points. Moreover, it does not require additional computation to interpolate feature representations between anchor points, making it more robust and efficient.

Table 4: Comparison of experimental results on the CUHK-SYSU and PRW datasets

Methods		CUHK-SYSU		PRW	
		mAP	Top-1	mAP	Top-1
One-step	OIM [44]	75.5	78.7	21.3	49.4
	IAN [45]	76.3	80.1	23.0	61.9
	NPSM [46]	77.9	81.2	24.2	53.1
	RCAA [44]	79.3	81.3	–	–
	CTXG [47]	84.1	86.5	33.4	73.6
	QEEPS [48]	88.9	89.1	37.1	76.7
	HOIM [34]	89.7	90.8	39.8	80.4
	BINet [49]	90.0	90.7	45.3	81.7
	NAE [50]	91.5	92.4	43.3	80.9
	NAE+ [50]	92.1	92.9	44.0	81.1
	AlignPS [4]	93.1	93.4	45.9	81.9
	AlignPS+ [4]	94.0	94.5	46.1	82.1
	ABOS [21]	93.7	94.3	46.4	84.9
	ABOS+ [21]	93.4	94.0	45.2	84.7
	DAAPS	**95.0**	**95.6**	**48.6**	84.7
Two-step	DPM+IDE [41]	–	–	20.5	48.3
	CNN+MGTS [51]	83.0	83.7	32.6	72.1
	CNN+CLSA [52]	87.2	88.5	38.7	65.0
	FPN+RDLR [53]	93.0	94.2	42.9	70.2
	IGPN [54]	90.3	91.4	47.2	87.0
	TCTS [43]	93.9	95.1	46.8	**87.5**

4.6 Ablation Studies of Loss Function

To prove the rationality of choosing the TOIM Loss function, the proposed model's performance is further evaluated using several different loss functions. As shown in Table 5, it is found that using the composite TOIM Loss function results in better performance of DAAPS than using only Logarithmic Loss Function or Triplet Loss function. Compared to applying the Triplet Loss function with Look-Up-Table (LUT), TOIM increases mAP and Top-1 by 1.7% and 1.8% on the CUHK-SYSU dataset, moreover 0.3% and 0.4% on the PRW dataset.

4.7 Ablation Studies of ResNeXt+

The study compares pedestrian detection models based on ResNet, ResNeXt, and ResNeXt+ network structures, excluding other factors, and the experimental results are shown in Table 6. The accuracy and overall performance of the model based on ResNeXt+ are higher than others. Although the improved metric values are small, the results cannot simply be compared with the ResNet-based model. Using the optimized ResNeXt+ overcomes the adverse effects of ResNeXt on the model and instead gives better results than ResNet.

Table 5: Comparison of DAAPS under different loss functions on the CUHK-SYSU and PRW

CUHK-SYSU	mAP	Top-1	ΔmAP	ΔTop-1
L_{OIM}	92.4	92.9	–	–
L_{tri} w/o LUT	92.8	93.2	+0.4	+0.3
L_{tri} w/ LUT	93.1	93.4	+0.7	+0.5
L_{TOIM}	**94.1**	**94.7**	**+1.7**	**+1.8**
PRW	mAP	Top-1	ΔmAP	ΔTop-1
L_{OIM}	45.7	81.8	–	–
L_{tri} w/o LUT	45.8	81.8	+0.1	+0
L_{tri} w/ LUT	45.9	81.9	+0.2	+0.1
L_{TOIM}	**46.0**	**82.2**	**+0.3**	**+0.4**

Table 6: Comparison of DAAPS based on different networks on the CUHK-SYSU

CUHK-SYSU	mAP	Top-1	ΔmAP	ΔTop-1
ResNet	94.0	94.5	–	–
ResNeXt	93.6	94.0	−0.4	−0.5
ResNeXt+	**94.1**	**94.7**	**+0.1**	**+0.2**

4.8 Visualization Results

The task of Person Search is challenging due to various factors that can affect people's posture, clothing, facial expression, and other characteristics in real-world scenes, making the detecting and recognizing of individuals difficult, especially in low-light environments and occluded areas.

Due to various factors (such as camera distance and resolution), the size of a pedestrian in a natural scene can vary greatly. Some pedestrians may appear at relatively large distances, causing them to appear smaller in size in the image, known as "small targets". For the performance evaluation of pedestrian detection algorithms, obtaining a high recall rate and accuracy on small targets is essential. The DAAPS addresses these issues by utilizing the Deformable Attention Mechanism to guide the model to focus on the salient features of the target person adaptively. This allows the model to identify small targets better and perform well in complex and occluded environments. The effectiveness of the DAAPS model is further demonstrated through visualization, and its results are shown in Fig. 5.

The paper chooses images with more occlusion, complex backgrounds, and dim light, which is a big challenge for the Person Search model. Enter the person to be detected on the left, ask the DAAPS model to find other pictures with the same target in the vast database and show the detection results in a blue box. The detection results show that the DAAPS model successfully recognizes and accurately locates the target to be queried, proving the proposed model's effectiveness.

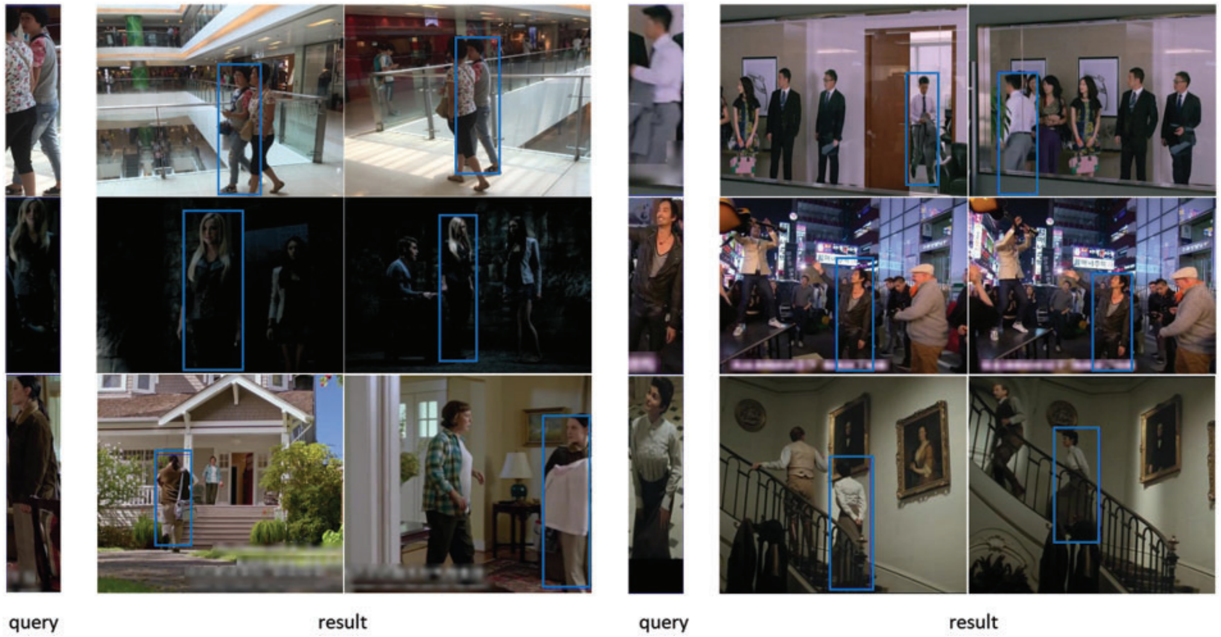

Figure 5: Visualization results

5 Conclusion

To reduce the impact of complex scenes and varying levels of occlusion on the model's accuracy in Person Search, this paper proposes the DAAPS model to combine Deformable Attention Mechanism with Anchor-free architecture for the first time. Separately, the detection network ResNeXt+ of DAAPS with enhanced scalability extracts multi-scale features for improved adaptability in complex Person Search tasks. Moreover, applying the more effective TOIM Loss function in the re-ID module improves the discriminative ability of embedding vectors. The model's generalization ability and robustness are enhanced, achieving better performance in practical applications demonstrated by simulation experiments, with mAP and Top-1 of 95.0% and 95.6% on the CUHK-SYSU dataset and 48.6% and 84.7% on the PRW dataset, respectively. The DAAPS model outperforms current models based on Anchor-free, showcasing rationality and effectiveness. In the study, many ablation experiments are used to test various essential modules of the model. The experimental results fully demonstrate the adaptability of the Deformable Attention mechanism as well as the rest of the components to the Anchor-free model, which offer a strong accuracy addition to the detector, and therefore provide ideas for later scholars to study the Anchor-free Person Search model. Due to the limitations of hardware, the model proposed in this paper does not achieve its optimal performance. In the future, more perfect algorithms and better hardware devices will be designed to enhance the real-time efficiency of the Person Search model.

Acknowledgement: Not applicable.

Funding Statement: We would like to express our sincere gratitude to the Natural Science Foundation of Shanghai under Grant 21ZR1426500, and the Top-Notch Innovative Talent Training Program for Graduate Students of Shanghai Maritime University under Grant 2021YBR008, for their generous support and funding through the project funding program. This funding has played a pivotal role in the successful completion of our research. We are deeply appreciative of their invaluable contribution to our research efforts.

Author Contributions: Study conception and design: X. Xin, D. Han; data collection: X. Xin; analysis and interpretation of results: X. Xin, D. Han, M. Cui; draft manuscript preparation: X. Xin, D. Han, M. Cui. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data provided in this study are available on request to the corresponding author by xinxiaoqi@stu.shmtu.edu.cn.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] T. Xiao, S. Li, B. Wang, L. Lin and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 3415–3424, 2017.

[2] M. Fiaz, H. Cholakkal, R. M. Anwer and F. S. Khan, "SAT: Scale-augmented Transformer for person search," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, Waikoloa, HI, USA, pp. 4820–4829, 2023.

[3] Y. Yan, J. Li, J. Qin, P. Zheng, S. Liao *et al.,* "Efficient person search: An anchor-free approach," *International Journal of Computer Vision*, vol. 131, pp. 1642–1661, 2023. https://doi.org/10.1007/s11263-023-01772-3

[4] Y. Yan, J. Qin, S. Bai, S. Liao, L. Liu *et al.,* "Anchor-free person search," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 7690–7699, 2021.

[5] D. Han, Y. Zhu, D. Li, W. Liang and K. C. Li, "A blockchain-based auditable access control system for private data in service-centric IoT environments," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3530–3540, 2022.

[6] H. Li, D. Han and M. Tang, "A privacy-preserving storage scheme for logistics data with assistance of blockchain," *IEEE Internet of Things Journal*, vol. 9, no. 6, pp. 4704–4720, 2022.

[7] D. Li, D. Han, Z. Zheng, H. Li, H. Liu *et al.,* "MOOCsChain: A blockchain-based secure storage and sharing scheme for MOOCs learning," *Computer Standards & Interfaces*, vol. 81, pp. 103597, 2022.

[8] D. Li, D. Han, T. H. Weng, Z. Zheng, H. Li *et al.,* "Blockchain for federated learning toward secure distributed machine learning systems: A systemic survey," *Soft Computing*, vol. 26, no. 9, pp. 4423–4440, 2022.

[9] Z. Tian, C. Shen, H. Chen and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Seoul, Korea (South), pp. 9627–9636, 2019.

[10] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*. Montreal, Canada: Curran Associates, Inc., 2015

[11] Z. Xia, X. Pan, S. Song, L. E. Li and G. Huang, "Vision transformer with deformable attention," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 4794–4803, 2022.

[12] K. Shaukat, S. Luo and V. Varadharajan, "A novel deep learning-based approach for malware detection," *Engineering Applications of Artificial Intelligence*, vol. 122, pp. 106030, 2023.

[13] K. Shaukat, S. Luo and V. Varadharajan, "A novel method for improving the robustness of deep learning-based malware detectors against adversarial attacks," *Engineering Applications of Artificial Intelligence*, vol. 116, pp. 105461, 2022.

[14] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, S. Chen *et al.,* "Performance comparison and current challenges of using machine learning techniques in cybersecurity," *Energies*, vol. 13, no. 10, pp. 2509, 2020.

[15] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed and M. Xu, "A survey on machine learning techniques for cyber security in the last decade," *IEEE Access*, vol. 8, pp. 222310–222354, 2020.

[16] W. Liu, S. Liao, W. Ren, W. Hu and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 5182–5191, 2019.

[17] K. Zhou, Y. Yang, A. Cavallaro and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Seoul, Korea (South), pp. 3702–3712, 2019.

[18] D. Chen, S. Zhang, W. Ouyang, J. Yang and Y. Tai, "Person search via a mask-guided two-stream CNN model," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 734–750, 2018.

[19] Z. He and L. Zhang, "End-to-end detection and re-identification integrated net for person search," In: C. V. Jawahar, H. Li, G. Mori, K. Schindler (Eds.), *Computer Vision–ACCV 2018*. Perth, Australia: Lecture Notes in Computer Science, pp. 349–364, 2019.

[20] C. Chi, F. Wei and H. Hu, "RelationNet++: Bridging visual representations for object detection via transformer decoder," in *Advances in Neural Information Processing Systems*, pp. 13564–13574, 2020.

[21] Y. Chen, D. Han, M. Cui, Z. Wu and C. C. Chang, "ABOS: An attention-based one-stage framework for person search," *EURASIP Journal on Wireless Communications and Networking*, vol. 2022, no. 1, pp. 75, 2022.

[22] M. Zhao, S. Gao, J. Ma and Z. Zhang, "Joint clothes image detection and search via anchor free framework," *Neural Networks*, vol. 155, pp. 84–94, 2022. https://doi.org/10.1016/j.neunet.2022.08.011

[23] X. Xiang, N. Lv and Y. Qiao, "Transformer-based person search model with symmetric online instance matching," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, pp. 2729–2733, 2022.

[24] C. Chen, D. Han and C. C. Chang, "CAAN: Context-aware attention network for visual question answering," *Pattern Recognition*, vol. 132, pp. 108980, 2022.

[25] D. Han, N. Pan and K. C. Li, "A traceable and revocable ciphertext-policy attribute-based encryption scheme based on privacy protection," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 1, pp. 316–327, 2022.

[26] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 734–750, 2018.

[27] J. Li, D. Han, Z. Wu, J. Wang, K. C. Li *et al.,* "A novel system for medical equipment supply chain traceability based on alliance chain and attribute and role access control," *Future Generation Computer Systems*, vol. 142, pp. 195–211, 2023.

[28] T. T. Feng and H. Y. Ge, "Pedestrian detection based on attention mechanism and feature enhancement with SSD," in *5th Int. Conf. on Communication, Image and Signal Processing (CCISP)*, Chengdu, China, pp. 145–148, 2020.

[29] S. Zhang, D. Chen, J. Yang and B. Schiele, "Guided attention in CNNs for occluded pedestrian detection and re-identification," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1875–1892, 2021.

[30] Z. Guo and D. Han, "Sparse co-attention visual question answering networks based on thresholds," *Applied Intelligence*, vol. 53, no. 1, pp. 586–600, 2023.

[31] X. Shen, D. Han, Z. Guo, C. Chen, J. Hua *et al.,* "Local self-attention in transformer for visual question answering," *Applied Intelligence*, vol. 53, pp. 16706–16723, 2023. https://doi.org/10.1007/s10489-022-04355-w

[32] Z. Guo and D. Han, "Multi-modal co-attention relation networks for visual question answering," *The Visual Computer*, vol. 37, 2022. https://doi.org/10.1007/s00371-022-02695-9

[33] C. Zhong, G. Qi, N. Mazur, S. Banerjee, D. Malaviya *et al.,* "A domain adaptive person re-identification based on dual attention mechanism and camstyle transfer," *Algorithms*, vol. 14, no. 12, pp. 361, 2021.

[34] K. Zheng, W. Liu, J. Liu, Z. J. Zha and T. Mei, "Hierarchical gumbel attention network for text-based person search," in *Proc. of the 28th ACM Int. Conf. on Multimedia, in MM'20*, New York, NY, USA, Association for Computing Machinery, pp. 3441–3449, 2020.

[35] Z. Ji, S. Li and Y. Pang, "Fusion-attention network for person search with free-form natural language," *Pattern Recognition Letters*, vol. 116, pp. 205–211, 2018. https://doi.org/10.1016/j.patrec.2018.10.020

[36] J. Cao, Y. Pang, R. M. Anwer, H. Cholakkal, J. Xie *et al.,* "PSTR: End-to-end one-step person search with transformers," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 9448–9457, 2022.

[37] L. Chen and J. Xu, "Sequential transformer for end-to-end person search," *arXiv:2211.04323*, 2022. https://doi.org/10.48550/arXiv.2211.04323

[38] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan *et al.,* "Multiscale vision transformers," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Montreal, Canada, pp. 6824–6835, 2021.

[39] S. Xie, R. Girshick, P. Dollar, Z. Tu and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 1492–1500, 2017.

[40] H. Li, A. Kadav, I. Durdanovic, H. Samet and H. P. Graf, "Pruning filters for efficient ConvNets," *arXiv:1608.08710*, 2017. https://doi.org/10.48550/arXiv.1608.08710

[41] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang *et al.,* "Person re-identification in the wild," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 1367–1376, 2017.

[42] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao *et al.,* "Dual attention network for scene segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 3146–3154, 2019.

[43] C. Wang, B. Ma, H. Chang, S. Shan and X. Chen, "TCTS: A task-consistent two-stage framework for person search," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, pp. 11949–11958, 2020.

[44] X. Chang, P. Y. Huang, Y. D. Shen, X. Liang, Y. Yang *et al.,* "RCAA: Relational context-aware agents for person search," In: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), *Lecture Notes in Computer Science*, vol. 11213, pp. 86–102, 2018.

[45] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei *et al.,* "IAN: The individual aggregation network for person search," *Pattern Recognition*, vol. 87, pp. 332–340, 2019.

[46] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao *et al.,* "Neural person search machines," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 493–501, 2017.

[47] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu *et al.,* "Learning context graph for person search," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 2158–2167, 2019.

[48] B. Munjal, S. Amin, F. Tombari and F. Galasso, "Query-guided end-to-end person search," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 811–820, 2019.

[49] W. Dong, Z. Zhang, C. Song and T. Tan, "Bi-directional interaction network for person search," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 2836–2845, 2020.

[50] D. Chen, S. Zhang, J. Yang and B. Schiele, "Norm-aware embedding for efficient person search," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 12612–12621, 2020.

[51] D. Chen, S. Zhang, W. Ouyang, J. Yang and Y. Tai, "Person search by separated modeling and a mask-guided two-stream CNN model," *IEEE Transactions on Image Processing*, vol. 29, pp. 4669–4682, 2020.

[52] X. Lan, X. Zhu and S. Gong, "Person search by multi-scale matching," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 536–552, 2018.

[53] C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang *et al.,* "Re-ID driven localization refinement for person search," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision* , Seoul, Korea (South), pp. 9814–9823, 2019.

[54] W. Dong, Z. Zhang, C. Song and T. Tan, "Instance guided proposal network for person search," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 2582–2591, 2020.