Anchor-Free Person Search

Yichao Yan1,*†, Jingpeng Li1*, Jie Qin†, Song Bai2, Shengcai Liao1, Li Liu1, Fan Zhu1, and Ling Shao1
1 Inception Institute of Artificial Intelligence (IIAI), UAE 2 University of Oxford, UK
{yanyichao91, ljpadam, qinjiebuaa, songbai.site}@gmail.com, {scliao, ling.shao}@ieee.org

Abstract

Person search aims to simultaneously localize and identify a query person from realistic, uncropped images, which can be regarded as the unified task of pedestrian detection and person re-identification (re-id). Most existing works employ two-stage detectors like Faster-RCNN, yielding encouraging accuracy but with high computational overhead. In this work, we present the Feature-Aligned Person Search Network (AlignPS), the first anchor-free framework to efficiently tackle this challenging task. AlignPS explicitly addresses the major challenges, which we summarize as the misalignment issues in different levels (i.e., scale, region, and task), when accommodating an anchor-free detector for this task. More specifically, we propose an aligned feature aggregation module to generate more discriminative and robust feature embeddings by following a “re-id first” principle. Such a simple design directly improves the baseline anchor-free model on CUHK-SYSU by more than 20% in mAP. Moreover, AlignPS outperforms state-of-the-art two-stage methods, with a higher speed. Code is available at: https://github.com/daodaofr/AlignPS

1. Introduction

Person search [54, 47], which aims to localize and identify a target person from a gallery of realistic, uncropped scene images, has recently emerged as a practical task with real-world applications. To tackle this task, we need to address two fundamental tasks in computer vision, i.e., pedestrian detection [33, 51] and person re-identification (re-id) [15, 1]. Both detection and re-id are very challenging tasks and have received tremendous attention in the past decade. In person search, we need to not only address the challenges (e.g., occlusions, pose/viewpoint variations, and background clutter) of the two individual tasks, but also pursue a unified and optimized framework to simultaneously perform detection and re-id.

Previous efforts devoted to this research topic can be generally divided into two categories. The first line of works [54, 5, 22], which we refer to as two-step approaches, attempt to deal with detection and re-id separately. As shown in Fig. 1a, multiple persons are first localized with off-the-shelf detection models, and then cropped out and fed to re-id networks to extract discriminative embeddings. Although two-step models can obtain satisfactory results, the disentangled treatment of the two tasks is time- and resource-consuming. In contrast, the second line of approaches [47, 26, 3, 32, 6] provide a one-step solution that
unifies detection and re-id in an end-to-end manner. As shown in Fig. 1b, one-step models first apply an ROI-Align layer to aggregate features in the detected bounding boxes. The features are then shared by detection and re-id; with an additional re-id loss, the simultaneous optimization of the two tasks becomes feasible. Since these models adopt two-stage detectors like Faster-RCNN [38], we refer to them as one-step two-stage models. However, these methods inevitably inherit the limitations of two-stage detectors, e.g., high computational complexity caused by dense anchors, and high sensitivity to the hyperparameters including the size, aspect ratio and number of anchor boxes, etc.

In contrast to two-stage detectors, anchor-free models exhibit unique advantages (e.g., simpler structure and higher speed), and have been actively studied in recent years [36, 23, 29, 14]. Inspired by this, an open question is naturally thrown at us - Is it possible to develop an anchor-free framework for person search? The answer is yes. However, this is a non-trivial task due to the following three misalignment issues. 1) Many anchor-free models learn multi-scale features using feature pyramid networks (FPNs) [24] to achieve scale invariance for object detection. However, this introduces the misalignment issue for re-id (i.e., scale misalignment), as a query person needs to be compared with all the people of various scales in the gallery set. 2) In the absence of operations like ROI-Align, anchor-free models cannot align the features for re-id and detection according to a specific region. Therefore, re-id embeddings must be directly learned from feature maps without explicit region alignment. 3) Person search can be intuitively formulated as a multi-task learning framework with detection and re-id as its sub-tasks. Hence, we need to find a better trade-off/alignment between the two tasks.

In this work, we present the first anchor-free framework for efficient person search, which we name the Feature-Aligned Person Search Network (AlignPS). Our model employs the typical architecture of anchor-free detection models, but with a carefully designed aligned feature aggregation (AFA) module. We follow a “re-id first” principle to explicitly address the above-mentioned challenges. More specifically, AFA reshapes some building blocks of FPN by exploiting the deformable convolution and feature fusion to overcome the issues of region and scale misalignment in re-id feature learning. We also optimize the training procedures of re-id and detection to place more emphasis on generating robust re-id embeddings (as shown in Fig. 1c). These simple yet effective designs successfully transform a classic anchor-free detector into a powerful and efficient person search framework, and allow the proposed model to outperform its anchor-based competitors.

In summary, our main contributions include:

- We propose the first one-step one-stage framework for efficient person search. The anchor-free solution will significantly foster future research in this direction.
- We design an AFA module that simultaneously addresses the issues of scale, region, and task misalignment to successfully accommodate an anchor-free detector for the task of person search.
- As an anchor-free one-stage framework, our model surprisingly outperforms state-of-the-art one-step two-stage models on two challenging person search benchmarks, while running at a higher speed.

2. Related Work

**Pedestrian Detection.** Pedestrian or object detection can be considered as a preliminary task of person search. Current deep learning-based detectors are generally categorized into one-stage and two-stage models, according to whether they employ a region proposal layer to generate object proposals. Alternatively, object detectors can also be categorized into anchor-based and anchor-free detectors, depending on whether they utilize anchor boxes to associate objects. One of the most representative two-stage anchor-based detectors is Faster-RCNN [38], which has been extended into numerous variants [10, 2, 34, 39]. Notably, some one-stage detectors [28, 25, 37, 52] also work with anchor boxes. Compared with the above models, one-stage anchor-free detectors [36, 23, 29, 56, 50, 42] have been attracting more and more attention recently due to their simple structures and efficient implementations. In this work, we develop our person search framework based on a classic one-stage anchor-free detector, thus making the whole framework simpler and faster.

**Person Re-identification.** Person re-id is also closely related to person search, aiming to learn identity embeddings from cropped person images. Traditional methods employed various handcrafted features [30, 15, 17] before the renaissance of deep learning. However, to pursue better performance, current re-id models are mostly based on deep learning. Some models employ structure/part information in the human body to learn more robust representations [40, 41, 31, 48], while others focus on learning better distance metrics [1, 20, 8, 9, 44]. As person re-id usually lacks large-scale training data, data augmentation [16, 27, 45, 55] also becomes popular for tackling this task. Compared with detection which aims to learn common features of pedestrians, re-id needs to focus more on fine-grained details and unique features of each identity. Therefore, we propose to follow the “re-id first” principle to raise the priority of the re-id task, resulting in more discriminative identity embeddings for more accurate person search.

**Person Search.** Existing person search frameworks can be divided into two-step and one-step models. Two-step models first perform pedestrian detection and subsequently crop the detected people for re-id. Zheng et al. [54] introduced the first two-step framework for person search and
evaluated the combinations of different detectors and re-id models. Since then, several models [5, 22, 18, 43] have followed this pipeline. In [47], Xiao et al. proposed the first one-step person search framework based on Faster-RCNN. Specifically, a joint framework enabling end-to-end training of detection and re-id was proposed by stacking a re-id embedding layer after the detection features and proposing the Online Instance Matching (OIM) loss. So far, a number of improvements [26, 46, 3, 49, 32, 12, 6] have been made based on this framework. In general, two-step models may achieve better performance, while one-step models have the advantages of simplicity and efficiency. However, there is still room for improving one-step methods due to the aforementioned shortcomings of the two-stage anchor-based detectors they usually adopt. In this work, we introduce the first anchor-free model to further improve the simplicity and efficiency of one-step models, without any sacrifice in accuracy.

3. Feature-Aligned Person Search Networks

In this section, we introduce the proposed anchor-free framework (i.e., AlignPS) for person search. Firstly, we give an overview of the network architecture. Secondly, the proposed AFA module is elaborated with the aim of mitigating different levels of misalignment issues when transforming an anchor-free detector into a superior person search framework. Finally, we present the designed loss function to obtain more discriminative features for person search.

3.1. Framework Overview

The basic framework of the proposed AlignPS is based on FCOS [42], one of the most popular one-stage anchor-free object detectors. Differently, we adhere to the “re-id first” principle to put emphasis on learning robust feature embeddings for the re-id subtask, which is crucial for enhancing the overall performance of person search.

As illustrated in Fig. 2, our model simultaneously localizes multiple people in the image and learns re-id embeddings for them. Specifically, an AFA module is developed to aggregate features from multi-level feature maps in the backbone network. To learn re-id embeddings, which is the key of our method, we directly take the flattened features from the output feature maps of AFA as the final embeddings, without any extra embedding layers. For detection, we employ the detection head from FCOS as the final embeddings, without any extra embedding layers. For detection, we employ the detection head from FCOS which is good enough for the detection subtask. The detection head consists of two branches, both of which contain four $3 \times 3$ conv layers. In the meantime, the first branch predicts regression offsets and centerness scores, while the second makes foreground/background classification. Finally, each location on the output feature map of AFA will be associated with a bounding box with classification and centerness scores, as well as a re-id feature embedding.

3.2. Aligned Feature Aggregation

Following FPN [24], we make use of different levels of feature maps to learn detection and re-id features. As the key of our framework, the proposed AFA performs three levels of alignment, beyond the original FPN, to make the output re-id features more discriminative.

Scale Alignment. The original FCOS model employs different levels of features to detect objects of different sizes. This significantly improves the detection performance since the overlapped ambiguous samples will be assigned to different layers. For the re-id task, however, the multi-level prediction could cause feature misalignment between different scales. In other words, when matching a person of different scales, re-id features are inconsistently taken from different levels of FPN. Furthermore, the people
in the gallery set are of various scales, which could eventually make the multi-level model fail to find correct matches for the query person. Therefore, in our framework, we only make predictions based on a single layer of AFA, which explicitly addresses the feature misalignment caused by scale variations. Specifically, we employ the \( \{C_3, C_4, C_5\} \) feature maps from the ResNet-50 backbone, and AFA sequentially outputs \( \{P_3, P_4, P_5\} \), with strides of 32, 16, and 8, respectively. We only learn features from \( \{P_3\} \), which is the largest output feature map, for both the detection and re-id subtasks, and \( \{P_3, P_2, P_1\} \) are no longer generated as in the original FPN. Although this design may slightly influence the detection performance, we will show in Sec. 4.3 that it achieves a good trade-off between the detection and re-id subtasks.

**Region Alignment.** On the output feature map of AFA, each location perceives the information from the whole input image based on a large receptive field. Due to the lack of the ROI-Align operation as in Faster-RCNN, it is difficult for our anchor-free framework to learn more accurate features within the pedestrian bounding boxes, and thus leading to the issue of region misalignment. The re-id subtask is even more sensitive to this issue as background features could greatly impact the discriminative capability of the learned features. In AlignPS, we address this issue from three perspectives. First, we replace the \( 1 \times 1 \) conv layers in the lateral connections with \( 3 \times 3 \) deformable conv layers. As the original lateral connections are designed to reduce the channels of feature maps, a \( 1 \times 1 \) conv is enough. In our design, moreover, the \( 3 \times 3 \) deformable conv enables the network to adaptively adjust the receptive field on the input feature maps, thus implicitly fulfilling region alignment. Second, we replace the “sum” operation in the top-down pathway with a “concatenation” operation, which can better aggregate multi-level features. Third, we again replace the \( 3 \times 3 \) conv with a \( 3 \times 3 \) deformable conv for the output layer of FPN, which further aligns the multi-level features to finally generate a more accurate feature map. The above three designs work seamlessly to address the region misalignment issue, and we notice that these simple designs are extremely effective when accommodating the basic anchor-free model for our person search task.

**Task Alignment.** Existing person search frameworks typically treat pedestrian detection as the primary task, i.e., re-id embeddings are just generated by stacking an additional layer after the detection features. A recent work [53] investigated a parallel structure by employing independent heads for the two tasks to achieve robust multiple object tracking results. In our task of person search, we find the inferior re-id features largely hinder the overall performance. Therefore, we opt for a different principle to align these two tasks by treating re-id as our primary task. Specifically, the output features of AFA are directly supervised with a re-id loss (which will be introduced in the following subsection), and then fed to the detection head. This “re-id first” design is based on two considerations. First, the detection subtask has been relatively well addressed by existing person search frameworks, which directly inherit the advantages from existing powerful detection frameworks. Therefore, learning discriminative re-id embeddings is our primary concern. As we discussed, re-id performance is more sensitive to region misalignment in an anchor-free framework. Therefore, it is desirable for the person search framework to be inclined towards the re-id subtask. We also show in our experiments that this design significantly improves the discriminative capability of the re-id embeddings, while having negligible impact on detection. Second, compared with “detection first” and parallel structures, the proposed “re-id first” structure does not require an extra layer to generate re-id embeddings, and is thus more efficient.

### 3.3. Triplet-Aided Online Instance Matching Loss

Existing works typically employ the OIM loss to supervise the training of the re-id subtask. Specifically, OIM stores the feature centers of all labeled identities in a lookup table (LUT), \( V \in \mathbb{R}^{D \times L} = \{v_1, ..., v_L\} \), which contains \( L \) feature vectors with \( D \) dimensions. Meanwhile, a circular queue \( U \in \mathbb{R}^{D \times Q} = \{u_1, ..., u_Q\} \) containing the features of \( Q \) unlabeled identities is maintained. At each iteration, given an input feature \( x \) with label \( i \), OIM computes the similarity between \( x \) and all the features in the LUT and circular queue by \( V^T x \) and \( Q^T x \), respectively. The probability of \( x \) belonging to the identity \( i \) is calculated as:

\[
p_i = \frac{\exp(v_i^T x)/\tau}{\sum_{j=1}^{L} \exp(v_j^T x)/\tau + \sum_{k=1}^{Q} \exp(u_k^T x)/\tau},
\]

where \( \tau = 0.1 \) is a hyperparameter that controls the softness of the probability distribution. The objective of OIM is to minimize the expected negative log-likelihood:

\[
\mathcal{L}_{OIM} = -E_x[\log p_i], \; t = 1, 2, ..., L.
\]

Although OIM effectively employs both labeled and unlabeled samples, we still observe two limitations. First, the
distances are only computed between the input features and the features stored in the lookup table and circular queue, while no comparisons are made between the input features. Second, the log-likelihood loss term does not give an explicit distance metric between feature pairs.

To improve OIM, we propose a specifically designed triplet loss. For each person in the input images, we employ the center sampling strategy as in [21]. As shown in Fig. 3, for each person, a set of features located around the person center are considered as positive samples. The objective is to pull the feature vectors from the same person close, and push the vectors from different people away. Meanwhile, the features from the labeled persons should be close to the corresponding features stored in the LUT, and away from the other features in the LUT.

More specifically, suppose we sample $S$ vectors from one person; we get $X_m = \{x_{m,1}, \ldots, x_{m,S}, v_m\}$ and $X_n = \{x_{n,1}, \ldots, x_{n,S}, v_n\}$ as the candidate feature sets for the persons with identity labels $m$ and $n$, respectively, where $x_{i,j}$ denotes the $j$-th feature of person $i$, and $v_i$ is the $i$-th feature in the LUT. Given $X_m$ and $X_n$, positive pairs can be sampled within each set, while negative pairs are sampled between the two sets. The triplet loss can be calculated as:

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

(3)

where $M$ denotes the distance margin, and $D_{\text{pos}}$ and $D_{\text{neg}}$ denote the Euclidean distances between the positive pair and the negative pair, respectively. Finally, the Triplet-aided OIM (TOIM) loss is the summation of these two terms:

$$L_{\text{TOIM}} = L_{\text{tri}} + L_{\text{OIM}}.$$  

(4)

4. Experiments

4.1. Datasets and Settings

CUHK-SYSU [47] is a large-scale person search dataset which contains 18,184 images, with 8,432 different identities and 96,143 annotated bounding boxes. The images come from two kinds of data sources (i.e., real street snaps and movies/TV), covering diverse scenes and including variations of viewpoints, lighting, resolutions, and occlusions. We utilize the standard training/test split, where the training set contains 5,532 identities and 11,206 images, and the test set contains 2,900 query persons and 6,978 images. This dataset also defines a set of protocols with gallery sizes ranging from 50 to 4,000. We report the results using the default gallery size of 100 unless otherwise specified.

PRW [54] was captured using six static cameras in a university campus. The images are sampled from the videos, which consist of 11,816 video frames in total. Person identities and bounding boxes are manually annotated, resulting in 932 labeled persons with 43,110 bounding boxes. The dataset is split into a training set of 5,704 images with 482 different identities, and a test set of 2,057 query persons and 6,112 images.

Evaluation Metric. We employ the mean average precision (mAP) and top-1 accuracy to evaluate the performance for person search. We also employ recall and average precision (AP) to measure the detection performance.

4.2. Implementation Details

We employ ResNet-50 [19] pretrained on ImageNet [11] as the backbone. We set the batch size to 4, and adopt the stochastic gradient descent (SGD) optimizer with weight decay of 0.0005. The initial learning rate is set to 0.001 and is reduced by a factor of 10 at epoch 16 and 22, with a total of 24 epochs. We use a warmup strategy for 300 steps. We employ a multi-scale training strategy, where the longer side of the image is randomly resized from 667 to 2000 during training, while zero padding is utilized to fit the images with different resolutions. For inference, we resize the test images to a fixed size of 1500×900. Following [4], we add a focal loss [25] to the original OIM loss. All the experiments are implemented based on PyTorch [35] and MMDetection [7], with an NVIDIA Tesla V100 GPU. It takes around 29 and 20 hours to finish training on CUHK-SYSU and PRW, respectively.

4.3. Analytical Results

Baseline. We directly add a re-id head in parallel with the detection head to the FCOS model and take it as our baseline. As shown in Fig. 4, each of the alignment strategies brings notable improvements to the baseline, and combining all of them yields >20% improvements in mAP.

Scale Alignment. To evaluate the effects of scale alignment, we employ feature maps from different levels of AFA and report the results in Table 1. Specifically, we evaluate the features from $P_3$, $P_4$, and $P_5$ with strides of 8, 16, and 32, respectively. As can be observed, features from the largest scale $P_3$ yield the best performance, due to the fact that they absorb different levels of features from AFA, pro-

![Figure 4: Comparative results on CUHK-SYSU with different alignment strategies, i.e., scale alignment (SA), region alignment (RA), and task alignment (TA).](image-url)
Re-id
91.8 92.2
92.0 92.5

(b) AlignPS
93.1 93.4
Recall
AP
mAP
top-1
90.7 91.0
92.7 93.1
90.6 90.8
Output
Feature
90.4 80.5
91.4 91.9
93.1 93.4
90.3
91.1 91.6
90.9

∼ output deformable conv improve the model by
The individual components of lateral deformable conv and
pose the framework is capable of learning adaptive receptive
layers of 8, 16, and 32, respectively.

Region Alignment. We conduct experiments with
different combinations of lateral deformable conv, output de-
formable conv and feature concatenation, and analyze how
different region alignment components influence the overall
performance. The results are reported in Table 2. Without
all these modules, the framework only achieves 83.7% in
top-1 accuracy, which is ∼10% lower than the full model.
The individual components of lateral deformable conv and
output deformable conv improve the model by ∼7% and
∼8%, respectively. Feature concatenation also brings ∼1% improvements. By combining two of the three com-
ponents, we observe consistent improvements. Finally, em-
ploying all the three modules yields 93.1% in mAP and
93.4% in top-1 accuracy, significantly boosting the perfor-
mance. These ablation studies thoroughly demonstrate the
effectiveness of the region alignment strategies.

To further illustrate how the deformable convolutions
work in our framework, we visualize the learned offsets of
the deformable filters in Fig. 5. We observe that the pro-
posed framework is capable of learning adaptive receptive
field according to the layout of the human body, and is-

to occlusion, crowding, and scale variations. We also
observe that the lateral deformable conv in C3 learns tighter
offsets around the body center, while the offsets in the C4
layer cover larger regions, which makes the two layers com-
plementary to each other.

Table 1: Comparative results on CUHK-SYSU by employ-
ing different levels of features. P3, P4, and P5 are the fea-
ture maps with strides of 8, 16, and 32, respectively.

| Methods | Detection Recall | Re-id mAP | top-1 |
|---------|------------------|-----------|-------|
| P3      | 90.3 81.2        | 93.1      | 93.4  |
| P4      | 87.5 78.7        | 92.7      | 93.1  |
| P5      | 79.0 71.7        | 89.3      | 89.5  |
| P3, P4  | 90.4 80.5        | 91.1      | 91.6  |
| P3, P4, P5 | 90.9 80.4  | 90.0      | 90.5  |

Table 2: Comparative results on CUHK-SYSU by employ-
ing different components in AFA for region alignment.

dconv” stands for deformable convolution.

| Lateral dconv | Output dconv | Feature concat | Re-id mAP | top-1 |
|---------------|--------------|----------------|-----------|-------|
| √             | √            |                | 83.4      | 83.7  |
| √             | √            |                | 90.6      | 90.8  |
| √             | √            |                | 91.4      | 91.9  |
| √             | √            |                | 84.0      | 84.1  |
| √             | √            |                | 91.8      | 92.2  |
| √             | √            |                | 90.7      | 91.0  |
| √             | √            |                | 92.0      | 92.5  |
| √             | √            |                | 93.1      | 93.4  |

(a) Deformable conv at lateral C3 layer in AFA

(b) Deformable conv at lateral C4 layer in AFA

Figure 5: Each image shows the sampling locations of two
levels of 3×3 (92 = 81 points at each location) deformable
filters: (a) Lateral deformable conv C3 + Output deformable
conv; (b) Lateral deformable conv C4 + Output deformable
conv. We illustrate different locations with different colors,
while center locations of people are marked in green. Please
zoom in for better visualization.

Figure 6: Illustration of different structures for training the
detection and re-id tasks.

Reg
Table 3: Comparative results on CUHK-SYSU with different training structures.

| Methods | Detection | Re-id | &nbsp; | &nbsp; |
|---------|-----------|-------|-------|-------|
| &nbsp;  | Recall    | AP    | mAP   | top-1 |
| $T_1$   | 87.5      | 79.0  | 80.3  | 79.2  |
| $T_2$   | 89.1      | 78.6  | 77.1  | 75.9  |
| $T_3$   | 90.1      | 81.4  | 80.7  | 80.2  |
| AlignPS | 90.3      | 81.2  | **93.1** | **93.4** |

Table 4: Comparative results on CUHK-SYSU with different loss functions.

| Methods | mAP | top-1 | $\Delta$ mAP | $\Delta$ top-1 |
|---------|-----|-------|---------------|----------------|
| OIM     | 92.4 | 92.9  | -             | -              |
| TOIM w/o LUT | 92.8 | 93.2  | +0.4          | +0.3           |
| TOIM w/ LUT | **93.1** | **93.4** | +0.7          | +0.5           |

Table 5: Comparative results on CUHK-SYSU with different deformable conv layers in the backbone model.

| Backbones | Deformable conv | mAP | top-1 |
|-----------|----------------|-----|-------|
| ResNet-50 | none           | 93.1 | 93.4 |
| ResNet-50 | res3           | 93.5 | 93.9 |
| ResNet-50 | res3 & res4    | 93.5 | 94.0 |
| ResNet-50 | res3 & res4 & res5 | **94.0** | **94.5** |

Task Alignment. Since person search aims to simultaneously address detection and re-id subtasks in a single framework, it is important to understand how different configurations of the two subtasks influence the overall task and which subtask should be paid more attention to. To this end, we design several structures to compare different training options (as shown in Fig. 6), the performance of which is summarized in Table 3. As can be observed, the structures of $T_1$ and $T_2$, where re-id features are shared with the regression and classification heads, respectively, yield significantly lower performance in re-id compared with our design. This indicates that the detection task takes advantage of the shared heads. As for $T_3$ where re-id and detection have independent feature heads, it achieves slightly better performance compared with $T_1$ and $T_2$, but still remarkably underperforms our design. These results indicate that our “re-id first” structure achieves the best task alignment among all these designs.

TOIM Loss. We evaluate the performance of our framework when adopting different loss functions and report the results in Table 4. We find that directly employing a triplet loss brings slight improvement. When employing the items in the LUT, the TOIM improves the mAP and top-1 accuracy by 0.7% and 0.5%, respectively. This indicates that it is beneficial to consider the relations between the input features and the features stored in the LUT.

Deformable Conv in the Backbone. As shown in Table 5, inserting deformable convolutions into the backbone network has positive effects on our framework. However, the contribution of the deformable conv layers in the backbone network is less significant than the deformable conv layers in our AFA module, e.g., only $\sim$1% improvement is observed with all the res3 & res4 & res5 deformable conv layers. These results indicate that the proposed AFA works as the key module for successful feature alignment.

4.4. Comparison to the State-of-the-Arts

We compare our model with the state-of-the-arts, including both one-step models [47, 46, 26, 3, 49, 32, 12, 6] and two-step models [5, 22, 18, 13, 43]. We denote our model with deformable conv layers in the backbone as AlignPS+. Results on CUHK-SYSU. As shown in Table 6, AlignPS/AlignPS+ outperforms all one-step models employing two-stage detection frameworks, which require region proposals and ROI-Align for inference. In contrast, our model is anchor-free and allows single-stage inference with a very simple structure, whilst running at a higher speed. Notably, AlignPS+ outperforms the current best-performing NAE+ [6] by 1.9% and 1.6% in mAP and top-1 accuracy, respectively. Also note that our model outperforms most two-step models, despite the fact that they employ two separate models for detection and re-id.

We visualize the results of AlignPS w.r.t. mAP with various gallery sizes and compare our model with both one-step and two-step models. Fig. 8 illustrates the detailed comparison results. As we can see, AlignPS outperforms all the
one-step models by notable margins, and is only inferior to the strongest two-step model TCTS [43], which requires an explicitly trained re-id model to adapt to the detection results. In contrast, our model does not need such a two-step process, as the alignment between the two subtasks is performed implicitly within the framework.

Results on PRW. PRW contains less training data; therefore, all the models achieve worse performance on this dataset. Nevertheless, as can be observed from Table 6, our model still outperforms all the one-step methods. We notice that BINet [12] also achieves strong performance on PRW. However, it requires an additional re-id branch to achieve region alignment during training, while our model efficiently addresses this issue with the AFA module.

Efficiency Comparison. Since different methods are evaluated with different GPUs, it is difficult to conduct a fair comparison of the efficiency of all the models. Here, we compare our method with OIM [47] and NAE/NAE+ [6] on the same Tesla V100 GPU. All the test images are resized to 1500×900 before being fed to the networks. As shown in Table 7, our anchor-free AlignPS only takes 61 milliseconds to process an image, which is 27% and 38% faster than NAE and NAE+, respectively. For query-guided models, e.g., IGPN [13] and QEEPS [32], they need to re-compute all the gallery features given each query. As AlignPS only computes the gallery features once, the total computation of these models can be thousands of times of AlignPS. It is also noteworthy that the parameters of all the two-step models are twice as our framework. These results clearly demonstrate the advantage of our anchor-free model in terms of computational efficiency.

Qualitative Results. Some qualitative results are illustrated in Fig. 7, where the query images come from movies/TV (left) and hand-held cameras (right). We can observe that our model successfully handles occlusions and scale/viewpoint variations, where OIM [47] and NAE [6] fail, demonstrating the robustness of our AlignPS.

5. Conclusion

In this paper, we propose the first anchor-free model to simplify the framework for person search, where detection and re-id are jointly addressed by a one-step model. We also design the aligned feature aggregation module to effectively address the scale, region, and task misalignment issues when accommodating an anchor-free detector for the person search task. Extensive experiments demonstrate that the proposed framework not only outperforms existing person search methods, but also runs at a higher speed.
References

[1] Ejaz Ahmed, Michael J. Jones, and Tim K. Marks. An improved deep learning architecture for person re-identification. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3908–3916, 2015. 1, 2

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: delving into high quality object detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 6154–6162, 2018. 2

[3] Xiaojun Chang, Po-Yao Huang, Yi-Dong Shen, Xiaodan Liang, Yi Yang, and Alexander G. Hauptmann. RCA: relational context-aware agents for person search. In Eur. Conf. Comput. Vis., pages 86–102, 2018. 1, 3, 7

[4] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Bernt Schiele. Hierarchical online instance matching for person search. In AAAI, pages 10518–10525, 2020. 5

[5] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream CNN model. In Eur. Conf. Comput. Vis., pages 764–781, 2018. 1, 3, 7

[6] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In IEEE Conf. Comput. Vis. Pattern Recog., pages 12612–12621, 2020. 1, 3, 7, 8

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiari Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Mmdetection: Open mmlab detection toolbox and benchmark. CoRR, abs/1906.07155, 2019. 5

[8] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1320–1329, 2017. 2

[9] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Hua Lai. Person re-identification by camera correlation aware feature augmentation. IEEE Trans. Pattern Anal. Mach. Intell., 40(2):392–408, 2018. 2

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In Int. Conf. Comput. Vis., pages 764–773, 2017. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In IEEE Conf. Comput. Vis. Pattern Recog., pages 248–255, 2009. 5

[12] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tie-niu Tan. Bi-directional interaction network for person search. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2836–2845, 2020. 3, 7, 8

[13] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tie-niu Tan. Instance guided proposal network for person search. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2582–2591, 2020. 7, 8

[14] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qiming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In Int. Conf. Comput. Vis., pages 6568–6577, 2019. 2

[15] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2360–2367, 2010. 1, 2

[16] Yixiao Ge, Zhuowu Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. FD-GAN: pose-guided feature distilling GAN for robust person re-identification. In Adv. Neural Inform. Process. Syst., pages 1230–1241, 2018. 2

[17] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In Eur. Conf. Comput. Vis., pages 262–275, 2008. 2

[18] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-id driven localization refinement for person search. In Int. Conf. Comput. Vis., pages 9813–9822, 2019. 3, 7

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE Conf. Comput. Vis. Pattern Recog., pages 770–778, 2016. 5

[20] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. CoRR, abs/1703.07737, 2017. 2

[21] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyond anchor-based object detection. IEEE Trans. Image Process., 29:7389–7398, 2020. 5

[22] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In Eur. Conf. Comput. Vis., volume 11205, pages 553–569, 2018. 1, 3, 7

[23] Hei Law and Jia Deng. Centernet: Detecting objects as paired keypoints. In Eur. Conf. Comput. Vis., pages 765–781, 2018. 2

[24] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 936–944, 2017. 2, 3

[25] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Int. Conf. Comput. Vis., pages 2999–3007, 2017. 2, 5

[26] Hao Liu, Jiashi Feng, Zequn Jie, Jayashree Karlekar, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. In Int. Conf. Comput. Vis., pages 493–501, 2017. 1, 3, 7

[27] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4099–4108, 2018. 2

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In Eur. Conf. Comput. Vis., pages 21–37, 2016. 2

[29] Wei Liu, Shengcai Liao, Weidong Hu, and Yichen Wei. Deformable convolutional networks. In Adv. Neural Inform. Process. Syst., pages 1320–1329, 2017. 2, 3

[30] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In Eur. Conf. Comput. Vis., volume 11205, pages 553–569, 2018. 1, 3, 7

[31] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In Eur. Conf. Comput. Vis., pages 765–781, 2018. 2

[32] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 936–944, 2017. 2, 3

[33] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Int. Conf. Comput. Vis., pages 2999–3007, 2017. 2, 5

[34] Hao Liu, Jiashi Feng, Zequn Jie, Jayashree Karlekar, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. In Int. Conf. Comput. Vis., pages 493–501, 2017. 1, 3, 7

[35] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4099–4108, 2018. 2

[36] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In Eur. Conf. Comput. Vis., pages 21–37, 2016. 2

[37] Wei Liu, Shengcai Liao, Weidong Hu, and Yichen Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 5187–5196, 2019. 2
[30] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.

[31] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Conf. Comput. Vis.*, pages 542–551, 2019.

[32] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 811–820, 2019.

[33] Wanli Ouyang and Xiaogang Wang. Joint deep learning for pedestrian detection. In *Conf. Comput. Vis.*, pages 2056–2063, 2013.

[34] Jianguo Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra R-CNN: towards balanced learning for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 821–830, 2019.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, pages 8024–8035, 2019.

[36] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 779–788, 2016.

[37] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6517–6525, 2017.

[38] Shaqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017.

[39] Guanglu Song, Yu Liu, and Xiaogang Wang. Revisiting the sibling head in object detector. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11560–11569, 2020.

[40] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *Int. Conf. Comput. Vis.*, pages 3980–3989, 2017.

[41] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline). In *Eur. Conf. Comput. Vis.*, pages 501–518, 2018.

[42] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *Int. Conf. Comput. Vis.*, pages 9626–9635, 2019.

[43] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. TCTS: A task-consistent two-stage framework for person search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11949–11958, 2020.

[44] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by discriminative selection in video ranking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(12):2501–2514, 2016.

[45] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 79–88, 2018.

[46] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. IAN: the individual aggregation network for person search. *Pattern Recognit.*, 87:332–340, 2019.

[47] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3376–3385, 2017.

[48] Y. Yan, J. Qin, B. Ni, J. Chen, L. Liu, F. Zhu, W. S. Zheng, X. Yang, and L. Shao. Learning multi-attention context graph for group-based re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[49] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2158–2167, 2019.

[50] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Int. Conf. Comput. Vis.*, pages 9656–9665, 2019.

[51] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4457–4465, 2017.

[52] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single-shot refinement neural network for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4203–4212, 2018.

[53] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *CoRR*, abs/2004.01888, 2020.

[54] Liang Zheng, Hengsheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3346–3355, 2017.

[55] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *Int. Conf. Comput. Vis.*, pages 3774–3782, 2017.

[56] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *CoRR*, abs/1904.07850, 2019.