

Robust Partial Matching for Person Search in the Wild

Yingji Zhong¹, Xiaoyu Wang^{2,3}, Shiliang Zhang¹

¹Peking University, ²Intellifusion, ³The Chinese University of Hong Kong, Shenzhen

¹{zhongyj, slzhang.jdl}@pku.edu.cn, ^{2,3}fanguhuaxue@gmail.com

Abstract

Various factors like occlusions, backgrounds, etc., would lead to misaligned detected bounding boxes, e.g., ones covering only portions of human body. This issue is common but overlooked by previous person search works. To alleviate this issue, this paper proposes an Align-to-Part Network (APNet) for person detection and re-Identification (reID). APNet refines detected bounding boxes to cover the estimated holistic body regions, from which discriminative part features can be extracted and aligned. Aligned part features naturally formulate reID as a partial feature matching procedure, where valid part features are selected for similarity computation, while part features on occluded or noisy regions are discarded. This design enhances the robustness of person search to real-world challenges with marginal computation overhead. This paper also contributes a Large-Scale dataset for Person Search in the wild (LSPS), which is by far the largest and the most challenging dataset for person search. Experiments show that APNet brings considerable performance improvement on LSPS. Meanwhile, it achieves competitive performance on existing person search benchmarks like CUHK-SYSU and PRW.

1. Introduction

Thanks to recent research efforts [27, 17, 10, 8, 25, 22, 30, 34, 31], the performance of person re-Identification (reID) has been significantly improved. However, one potential issue with current reID setting is that, it treats person detection and reID as two isolated steps. This issue can be compensated by person search [24, 33, 1, 28, 3, 15] to jointly accomplish person detection and reID. Being able to jointly optimize person detection and reID, person search is attracting more and more attention, and is potential to present advantages in flexibility, efficiency, and accuracy.

Compared with reID, person search needs to design pedestrian detectors robust to variances of scales, illuminations, backgrounds, occlusions, etc. Some person search works refine detectors by jointly training detectors and reID models [24, 15, 4, 3]. Although strong detectors are gen-

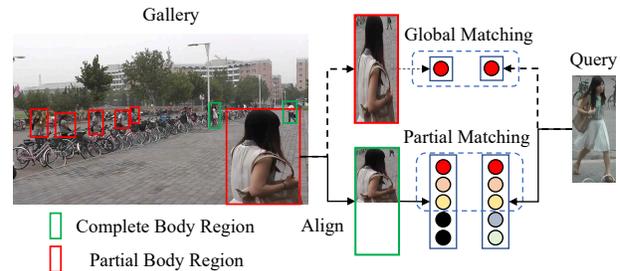


Figure 1. Illustration of misaligned person bounding boxes and issues of person matching with global feature. Misalignments commonly occur in person search and can not be eliminated by training stronger detectors. This paper addresses this issue with bounding box alignment and partial matching.

erally helpful for reID, they may still produce misaligned person bounding boxes. As shown in Fig. 1, occlusions and limited camera viewing field lead to many accurate but misaligned bounding boxes covering portions of pedestrians. Most of existing person search methods extracts global feature from the detected bounding boxes, no matter the joint models [24, 15] or the separate models [1, 28, 3]. As shown in Fig. 1, misalignment degrades the performance of global features, because it is not reasonable to match partial features against global features. More detailed reviews to person search works can be found in Sec. 2.

This paper targets to design a unified person search framework robust to the misaligned bounding boxes. As shown in Fig. 1, the basic idea is to refine detected bounding boxes to cover the estimated holistic body regions to eliminate misalignment errors. Then, aligned part features could be extracted from refined bounding boxes with simple region division. Part features hence allow for robust partial matching across bounding boxes, where features on the mutual visible regions are matched for similarity computation, and features on noisy regions are not considered for matching. In other words, our strategy improves both the feature resolution and robustness to misalignment errors.

The above idea leads to the Align-to-Part Network (APNet) for person search. Built on OIM [24], APNet consists of the additional Bounding Box Aligner (BBA) and Region

Sensitive Feature Extractor (RSFE). The BBA module is implemented by predicting 4 bounding box offset values. It can be trained by automatic data augmentation without requiring manual annotations. Because of enlarged receptive fields of neurons in feature extraction CNN layer, noisy or occluded parts would affect features of their adjacent parts in the same bounding box. We further design the RSFE module to extract part features. RSFE reinforces local cues in each part feature, hence effectively alleviates the negative effects of adjacent noises in part feature learning.

We test our approach on current person search benchmarks, *i.e.*, PRW [33] and CUHK-SYSU [24]. Experiment results show that, our APNet achieves competitive performance compared with recent works. To test our approach in more challenging scenarios, we contribute a new person search dataset named as Large-Scale dataset for Person Search in the wild (LSPS). Compared with PRW and CUHK-SYSU, LSPS presents several new features that would encourage the research towards more realistic person search: 1) large scale: it contains a large number of pedestrians and bounding boxes; 2) automatically detected bounding boxes, which differ with the manually annotated ones in PRW and CUHK-SYSU; 3) collected in scenarios with occlusions and crowdings. On this challenging LSPS, our method exhibits substantial advantages.

Compared with reID, person search is still relatively under-explored. It is critical to deal with the challenge occurring in both person detection and reID. To the best of our knowledge, the proposed APNet is an early attempt in boosting the robustness to misalignment issue in person search. As shown in our experiments, the APNet brings considerable performance gains. The proposed LSPS dataset simulates realistic scenarios and presents a more challenging person search task than existing datasets. We believe that, the proposed methods and the contributed LSPS can inspire more research efforts in person search.

2. Related Work

This paper is closely related to person reID and person search. The following parts review those two lines of works.

Person reID extracts robust and discriminative features from given person bounding boxes. Current fully supervised reID methods can be briefly summarized into the following categories. 1) Learn local features of regions utilizing offline detector to split body into several regions [27, 17, 10, 11], or uniformly split the body into several stripes [31, 22, 26]; 2) Learn the robust feature utilizing additional attribute annotation [19, 18, 20, 23]; 3) Enhance the discriminativeness of feature by attention mechanism [12, 9, 25]; 4) Impose constraints on feature space with loss functions, like verification loss [36] and triplet loss [8]. Among the above methods, PCB [22] and its variants [31, 26] dominate the state-of-the-art performance.

PCB uniformly splits the feature map into stripes and supervises them to ensure their discriminative power. However, PCB requires a strict alignment of the input images since misalignment breaks the correspondence of stripes on the same spatial position. The performance of PCB thus degrades substantially when misalignments exist [21].

There are some reID works focusing on partial reID. Zheng *et al.* [35] propose AMC+SWM to solve the partial reID. AMC collects patches from all gallery images and set up a dictionary for patch level matching. He *et al.* [7] propose DSR to reconstruct the partial image by holistic image and utilize the reconstruction error as the similarity. Both AMC [35] and DSR [7] are optimisation-based thus it is expensive to get the similarity of each query-gallery pair. PGFA [14] applies partial matching scheme with the aligned input images, whose occluded parts are indicated by the expensive offline keypoint detector. Sun *et al.* [21] propose a VPM to extract local features while simultaneously being aware of the visibility of each body part. However, VPM does not consider the deformation mentioned in DSR. Our method address the deformation by an explicit alignment step and gets rid of the time-consuming comparison of each query-gallery pair during the inference stage.

Person search considers the detection stage in raw video frames before reID. Current methods of person search can be divided into two categories. One integrates detection and reID into a unified framework [24, 15, 4, 29]. Based on Faster-RCNN [16], OIM proposed by Xiao *et al.* [24] inserts an additional feature extraction branch on top of the detector head, which is the first end-to-end learning framework for person search. Similar to NPSM of Liu *et al.* [4], Munjal *et al.* [15] apply a query-guided method for person search while it is end-to-end optimized compared to NPSM. The other category solves detection and re-identification with two separate models, *i.e.*, performing person search in a sequential manner [1, 28, 3]. Chen *et al.* [1] propose a mask-guided feature learning method to make the reID network focus more on foregrounds. Lan *et al.* [28] propose CLSA which utilizes multi-level features from reID network to solve the multi-scale matching problem. Han *et al.* [3] claim that bounding box from detector might not be optimal for re-identification, they thus refine the detector results driven by re-identification loss.

Our work belongs to the first category of person search, integrating detection and reID into a unified framework. Different from existing person search works, this work further addresses the misalignment issue of detected bounding boxes and optimizes aligned part feature learning to achieve robust partial matching in person search.

3. Problem Formulation

Given a query person image q and a gallery set $\mathcal{G} = \{g_i\}_{i=1:N}$ containing N frames, person search aims at

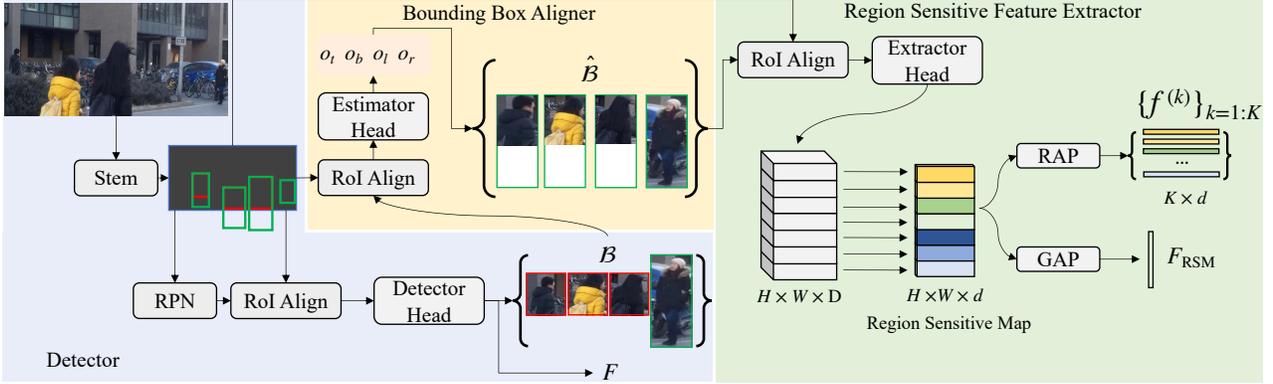


Figure 2. Architecture of the proposed Align-to-Part Network (APNet), which consists of person detector, Bounding Box Aligner (BBA), and Region Sensitive Feature Extractor (RSFE). The detection branch returns bounding boxes \mathcal{B} and global feature F . BBA refines detected boxes \mathcal{B} to aligned ones, *i.e.*, $\hat{\mathcal{B}}$. Red and green bounding boxes denotes detected boxes \mathcal{B} and refined boxes $\hat{\mathcal{B}}$, respectively. RSFE extracts part features $\{f^{(k)}\}_{k=1:K}$ from refined boxes and alleviates negative effects of adjacent noises. RAP and GAP denote Regional Average Pooling and Global Average Pooling, respectively. F_{RSM} is the global feature from Region Sensitive Map, which is only used for training RSFE.

detecting a collection of person bounding boxes $\mathcal{B} = \{b_i\}_{i=1:M}$ from \mathcal{G} , then matching q against the bounding boxes in \mathcal{B} . Suppose a person can be divided into K body parts, we could represent the bounding box b_i containing a complete person as a collection of K parts, *i.e.*, $\mathcal{P}_i = \{p_i^{(k)}\}_{k=1:K}$, where $p_i^{(k)}$ denotes the k -th part.

With the above formulation, person detection is expected to return bounding boxes containing complete persons with K parts. Person reID targets to extract a discriminative feature to identify the same identity of q in \mathcal{B} . Most of existing works extract a global feature F and performs reID by L2-distance. Due to the occlusions or background clutters, certain bounding boxes only contain portions of body parts, making the global feature degrades to a partial feature, *i.e.*, the global feature is extracted from l , $l < K$ visible parts. Such partial features lead to inaccurate person matching when compared with global features.

Our solution is to introduce a part identifier $E(\cdot)$ to identify visible parts in each detected bounding box, *i.e.*, $\mathcal{P}_i = E(b_i)$. With identified parts, the person image matching could be treated as a part feature matching task, where features on mutual visible parts of two bounding boxes are matched. The distance between q and b_i can be denoted as, *i.e.*,

$$\text{dist}_P(q, b_i) = \frac{\sum_{k \in \mathcal{P}_q \cap \mathcal{P}_i} D(f_q^{(k)}, f_i^{(k)})}{|\mathcal{P}_q \cap \mathcal{P}_i|}, \quad (1)$$

where $f_q^{(k)}$ and $f_i^{(k)}$ are features extracted from k -th part from q and b_i , and $D(\cdot)$ refers the L2-distance.

The training of our person search model should guarantee an accurate person detector, the reliable part identifier, and discriminative part features. We formulate our training

objective as,

$$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_P + \sum_{k=1}^K \mathcal{L}_{f^{(k)}}, \quad (2)$$

where \mathcal{L}_D denotes loss of the detector, which is optimized with both bounding box locations and person reID. \mathcal{L}_P denotes the loss of part identification. $\mathcal{L}_{f^{(k)}}$ evaluates the discriminative power of the k -th part feature, which can be implemented with reID loss. The following section presents details of our implementation to person detector, part identifier, part feature extraction, as well as the network optimization.

4. Proposed Methods

We propose the Align-to-Part Network (APNet) to implement the formulation in Sec. 3. The architecture of APNet is shown in Fig. 2. APNet consists of a person detector, Bounding Box Aligner (BBA), and the Region Sensitive Feature Extractor (RSFE), respectively. The following parts present details of those components.

The detector is built upon OIM [24], which is an end-to-end person detector returning bounding boxes \mathcal{B} as well as their corresponding global feature F . As shown in Fig. 2, the detector is trained with RPN loss ($\mathcal{L}_{\text{rpncls}}$, $\mathcal{L}_{\text{rpnreg}}$) [16], ROI Head loss (\mathcal{L}_{cls} , \mathcal{L}_{reg}), as well as reID loss. We denote the detector training loss as

$$\mathcal{L}_D = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{rpncls}} + \mathcal{L}_{\text{rpnreg}} + \mathcal{L}_{\text{ID}}, \quad (3)$$

where \mathcal{L}_{cls} and \mathcal{L}_{reg} denotes the person classification loss and bounding box regression loss in ROI Head. $\mathcal{L}_{\text{rpncls}}$ and $\mathcal{L}_{\text{rpnreg}}$ denote the objectiveness loss and proposal regression loss in RPN. \mathcal{L}_{ID} is the reID loss computed on the

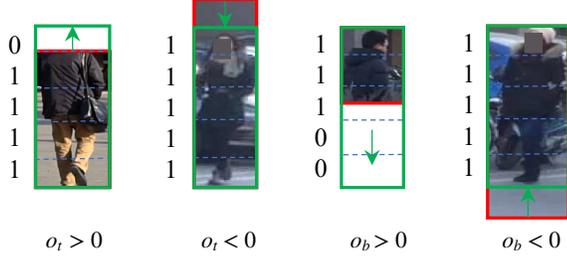


Figure 3. Illustration of BBA for bounding box refinement. Red and green boxes refer to detected boxes and refined boxes, respectively. BBA predicts four bounding box offset values o_t , o_b , o_l , and o_r , which shift the detected bounding boxes to cover the estimated holistic body region. 0, 1 indicate the validity of each part.

global feature. We refer readers to OIM [24] for more details of the loss computation.

As shown in Fig. 1, a well trained detector may produce misaligned person bounding boxes. We hence design a part estimator with BBA to estimate visible body parts of each detected box.

4.1. Bounding Box Aligner

BBA implements part identifier $E(\cdot)$ in Sec. 3 to identify visible parts in each bounding box. This could be achieved with various methods, *e.g.*, through segmenting person foregrounds [13] or estimating body keypoints [2]. However, those methods require extra annotations and considerable computations. We implement BBA with a more efficient solution.

Because most pedestrians show upright posture in surveillance videos, aligned body parts can be extracted by dividing a holistic body region into horizontal and vertical stripes. This operation generates aligned part regions, *e.g.*, top and bottom horizontal stripes correspond to head and foot, respectively. With this intuition, BBA first refines detected bounding boxes, then extracts horizontal and vertical stripes as body parts. We illustrate this procedure in Fig. 3.

BBA estimates an offset vector $O = \{o_t, o_b, o_l, o_r\}$ to refine each detected bounding box to cover the holistic body region, where each offset value is in the range of $[-1, 1]$. The four offset values move the top, bottom, left, and right boundaries of each bounding box. Fig. 3 shows examples of moving the boundaries by o_t and o_b . Suppose the coordinate of a bounding box b is $\{x_{min}, y_{min}, x_{max}, y_{max}\}$, its coordinate after refinement with O can be denoted as $\{\hat{x}_{min}, \hat{y}_{min}, \hat{x}_{max}, \hat{y}_{max}\}$, *e.g.*,

$$\begin{aligned} \hat{y}_{min} &= y_{min} - \frac{h \cdot o_t}{1 - o_t - o_b} \\ \hat{y}_{max} &= y_{max} + \frac{h \cdot o_b}{1 - o_t - o_b}, \end{aligned} \quad (4)$$

where h is the height of detected bounding box, computed

as $y_{max} - y_{min}$. Similar computation can be applied to compute the \hat{x}_{min} and \hat{x}_{max} with o_l and o_r .

We denote the refined bounding box as \hat{b} , which is used to extracted horizontal and vertical stripes as parts. As shown in Fig. 3, refinements may introduce occluded parts and noises into \hat{b} . To extract K horizontal stripes, we introduce a K -dim validity vector v to record the visibility of each stripe. The k -th stripe is considered as valid, *i.e.*, $v[k] = 1$, if

$$[K \cdot \max(0, o_t)] \leq k \leq K - \lfloor K \cdot \max(0, o_b) \rfloor. \quad (5)$$

Similar computation is applied to extract valid vertical stripes. The final valid part collection \mathcal{P} of each bounding box collects valid stripes.

As shown in Fig. 2, BBA predicts O based on the bounding box feature extracted with ROIAlign. BBA can be trained by automatically generating training data. We first crop bounding boxes from frames by their groundtruth coordinates, which we denote as \mathcal{B}_{gt} . AlphaPose [2] is utilized to estimate the keypoints of each bounding box, which provide cues about missing body parts. We then transform \mathcal{B}_{gt} to $\hat{\mathcal{B}}_{gt}$ to cover the holistic body region. Comparing \mathcal{B}_{gt} and $\hat{\mathcal{B}}_{gt}$ generates the groundtruth offset labels $O_{gt} = \{o_{gt-t}, o_{gt-b}, o_{gt-l}, o_{gt-r}\}$. The training of BBA can thus be supervised by following loss

$$\mathcal{L}_P = \sum_{i \in \{t, b, l, r\}} \text{smooth}_{l_1}(o_i, o_{gt-i}), \quad (6)$$

where smooth_{l_1} computes differences between predicted offset values and the groundtruth values. More details of smooth_{l_1} can be found in [16].

4.2. Region Sensitive Feature Extractor

The part collection \mathcal{P} makes it possible to extract part features for partial matching. The following section shows our method to extract horizontal stripe features. Vertical stripe features can be extracted with similar way.

Vanilla Part Feature Extractor: Part features can be extracted by applying Region Average Pooling (RAP) on the feature map of video frame. As shown in Fig. 2, for a refined bounding box \hat{b}_i , we first extract its feature map from the frame feature map \mathcal{M} with ROIAlign, which is then input into a convolutional block to generate a feature map $T \in \mathbb{R}^{H \times W \times D}$. We denote part features of \hat{b}_i extracted by RAP as,

$$\{\bar{f}_i^{(k)}\}_{k=1:l} = \text{RAP}(T, \mathcal{P}_i), l = |\mathcal{P}_i|, \quad (7)$$

where l denotes the number of valid horizontal stripes of bounding box \hat{b}_i .

Part feature learning can be achieved by computing the reID loss on each valid part feature, *i.e.*,

$$\mathcal{L}_f^{(k)} = \mathcal{L}_{\text{ID}}(\bar{f}_i^{(k)}, y), \quad (8)$$

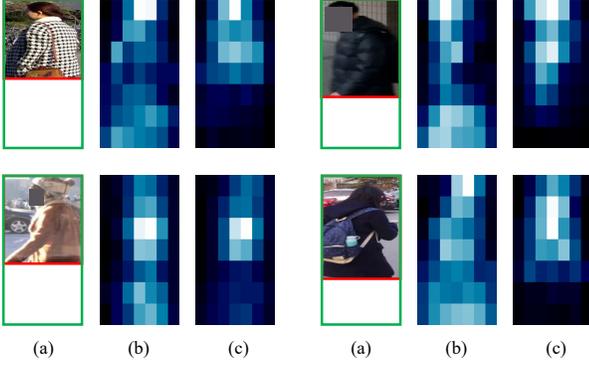


Figure 4. Illustration of detected boxes (red) and refined boxes (green), as well as their feature maps. (b) and (c) show feature maps of refined boxes trained with vanilla part feature extractor and our RSFE. It is clear that, RSFE depresses noises on noisy and invalid regions.

where \mathcal{L}_{ID} refers to reID loss which is implemented with OIM loss [24]. y is the ground truth person ID label.

Fig. 4 (b) shows feature maps of aligned bounding boxes trained by the above feature extractor. It is clear that, noisy or invisible regions still have strong responses. This could be because the training procedure in Eq. (8) focuses on visible body parts, and can not tune the features on noisy parts. Because of enlarged receptive fields of neurons in feature extraction CNN layer, the strong CNN activations on noisy or occluded parts would affect features of their adjacent valid parts. This may degrades the validity of the above feature extractors. Therefore, part feature extractors robust to occlusions and noises are required.

Feature Extraction with Region Sensitive Map: Our RSFE introduces a Region Sensitive Map (RSM) to suppress responses of invalid regions. To tune the responses of each feature stripe, we generate the RSM $M \in \mathbb{R}^{H \times W \times d}$ by applying 1×1 convolution on each horizontal feature stripe $T[i] \in \mathbb{R}^{W \times d}$, $i = 1 : H$. The computation of $M[i] \in \mathbb{R}^{W \times d}$ can be denoted as,

$$M[i] = \text{conv}_{1 \times 1}^{(i)}(T[i]), i = 1 : H, \quad (9)$$

where $\text{conv}_{1 \times 1}^{(i)}$ refers to the i -th convolution kernel. Different from a global 1×1 convolution, H convolutional kernels in Eq. (9) do not share parameters to deal with complicated occlusions on different spatial locations.

To suppress noisy regions, we supervise M with a global feature. The loss can be computed by applying Average Pooling on M and supervise the resulted feature F_{RSM} , *i.e.*,

$$\mathcal{L}_{RSM} = \mathcal{L}_{ID}(F_{RSM}, y). \quad (10)$$

This loss enforces the network to depress activations on noisy regions to generate a better F_{RSM} , which is then used for stripe feature extraction. Fig. 4 (c) shows feature maps

after training with Eq. (10), where invalid regions are effectively suppressed.

From M , the stripe features can be extracted by RAP as,

$$\{f_i^{(k)}\}_{k=1:l} = \text{RAP}(M, \mathcal{P}_i), l = |\mathcal{P}_i|, \quad (11)$$

where each part feature can be trained with part feature loss in Eq. (7).

Eq. (9) applies different 1×1 convolution kernels on spatial locations of T . This enables more specific refinements for each part feature. Besides that, different spatial locations show varied probabilities of occlusion, *e.g.*, the foot area is more likely to be occluded. This property makes Eq. (9) more effective in depressing occlusions and noises than learning a global 1×1 convolution kernel.

The methods above extracts features from horizontal stripes. Similar procedure can be applied to extract features from vertical stripes. Given a query person image q and a detected gallery bounding boxes b_i to be compared, we utilize both global and part features for person reID. The part feature distance can be computed with Eq. (1). We also utilize the global feature F from detector branch for distance computation. The overall distance between q and b_i can be computed as:

$$\text{dist}(q, b_i) = \text{dist}_P(q, b_i) + \lambda \cdot D(F_q, F_i), \quad (12)$$

where F_q and F_i correspond to the global features of q and b_i respectively. We set λ to 1 in the following experiments.

5. LSPS Dataset

5.1. Previous Datasets

CUHK-SYSU [24] consists of 5,694 frames captured from movie snapshots, and 12,490 frames from street snap. Both bounding boxes and identities are labeled manually. The dataset provides 8,432 labeled identities, and 23,430 labeled bounding boxes. 96,143 bounding boxes are provided in total. 11,206 frames with 5,532 identities make up the training set. CUHK-SYSU does not annotate bounding boxes with partial bodies.

PRW [33] is captured with six cameras deployed at a campus, containing 11,816 frames with 932 identities in total. 5,134 frames are selected as training set with 432 identities, while the rest 6,112 frames make up the test set. A total of 34,304 bounding boxes are annotated with identity. Similar to CUHK-SYSU, bounding box locations in PRW are also manually labeled. PRW includes some misaligned bounding boxes in both queries and galleries.

5.2. Description to LSPS

This paper contributes a new *Large-Scale dataset for Person Search in the wild* (LSPS), which shows the following characteristics:

Dataset	LSPS	PRW [33]	CUHK [24]
frames	51,836	11,816	18,184
identities	4,067	932	8,432
anno. boxes	60,433	34,304	23,430
cameras	17	6	-
detector	Faster-RCNN	Hand	Hand
inc.query	~60%	~6%	~0%

Table 1. Comparison between LSPS and the other two person search datasets. “detector” refers to the way to obtain the ground truth of bounding box location. “inc.query” means the percentage of query bounding boxes with partial body.

Complex scene and appearance variations: The video frames are collected from 17 cameras, deployed at both indoor and outdoor scenes. Different cameras exhibit different backgrounds, viewpoints, view fields, illuminations, pedestrian densities, *etc.* Those factors lead to substantial appearance variances for the same person. Meanwhile, due to limited view fields of each camera and high person density, lots of pedestrians are occluded, leading to bounding boxes covering partial body region. LSPS includes partial bounding boxes into both query and gallery sets. Fig. 5 (a) compares the body completeness in queries between LSPS and PRW. It is clear that, LSPS has a substantially larger number of incomplete query bounding boxes. Compared with existing person search datasets, LSPS presents more complex scenes and appearance variations.

Larger scale: Different with previous benchmarks, where bounding boxes are manually labeled. LSPS utilizes bounding boxes detected by Faster-RCNN [16], based on which we manage to collect a dataset with larger number of bounding boxes. We show the comparison between LSPS and the other two datasets in Table 1. In LSPS, a total number of 51,836 frames are collected, from which 60,433 bounding boxes and 4,067 identities are annotated. Table 1 shows that, LSPS presents the larger number of frames, annotated person bounding boxes, cameras, respectively. Moreover, about 60% queries in LSPS cover partial body. Larger scale and incomplete query bounding boxes make LSPS a more challenging and realistic dataset than the others in Table 1.

5.3. Evaluation Protocol

LSPS is splitted into a training set with 18,154 frames and a test set with 33,673 frames. The training and test set contain 1,041 and 3,026 identities, respectively. Statistics of the training and test set are summarized in Table 2. We utilize mean Average Precision (mAP) and rank-1 accuracy as evaluation metrics, which are widely adopted in person reID [24, 15, 1, 28, 3]. In person search, a retrieved bounding box is considered as true positive if it shares the same ID label with query and it has the overlap rate larger than 0.5 with the ground truth bounding box. Therefore, both

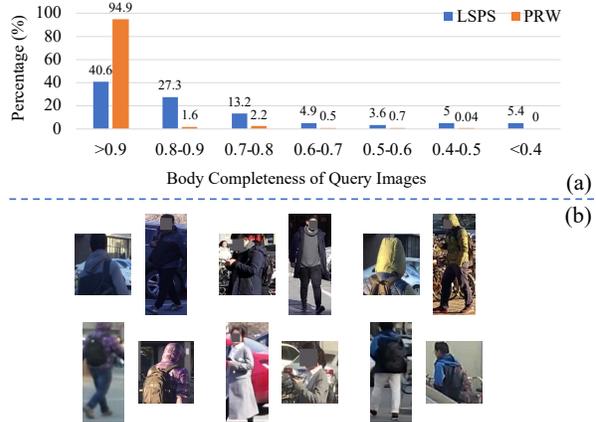


Figure 5. (a) compares the body completeness in queries between LSPS and PRW. (b) illustrates several query and gallery bounding boxes. It can be observed that, misalignment occurs in both query and gallery bounding boxes.

Split	frames	identities	boxes	anno. boxes
Training	18,163	1,041	71,563	18,928
Test	33,673	3,026	116,170	41,505

Table 2. Statistics of training/test sets on LSPS. “anno. boxes” refers bounding boxes annotated with person ID. “boxes” denotes the number of provided bounding boxes.

mAP and rank-1 accuracy in person search are affected by the detector performance.

6. Experiment

The following parts conduct experiments on CUHK-SYSU [24], PRW [33] and the new LSPS.

6.1. Implementation Details

Our APNet is implemented based on OIM [24]. We use the ResNet50 [6] initialized with ImageNet pretrained model as the backbone. The backbone adopts similar setting with the one in [15] and it uses RoIAlign [5] rather than RoIPooling [16]. For all experiments, we train the APNet with SGD optimizer. We divide the training into two stages, *i.e.*, the first stage trains the detector branch with \mathcal{L}_D . The second stage fixes parameters of the detector and trains BBA and RSFE with \mathcal{L}_P and \mathcal{L}_f , respectively. We set K to 7 for horizontal stripes. For PRW and CUHK-SYSU, we do not use features from vertical stripes since the misalignment in these two datasets mainly occurs in vertical direction.

On CUHK-SYSU, the first stage lasts 40k iterations. We set the initial learning rate to $1e-3$ and decay it by 0.1 at 30k. RSFE is trained for 40k with learning rate as $1e-3$, decayed at 30k. On PRW, the first stage lasts 80k. We set learning rate to $1e-4$, decayed at 60k. RSFE is trained for 40k with learning rate as $1e-3$, decayed at 30k. On LSPS, the first

Methods		mAP	rank-1
	feature		
upper bound	global	53.5	74.6
global matching	global	13.6	30.4
	stripe	23.1	52.5
partial matching	valid region	23.9	46.7
	mutual region	21.4	48.6
	mutual stripe	30.9	70.0

Table 3. Validity of partial matching on Market1501 modified by adding artificial occlusions. “upper bound” refers to performance of global feature on original Market-1501. Other methods are tested on modified Market-1501. Valid region refers to common visible region of tow matched bounding boxes.

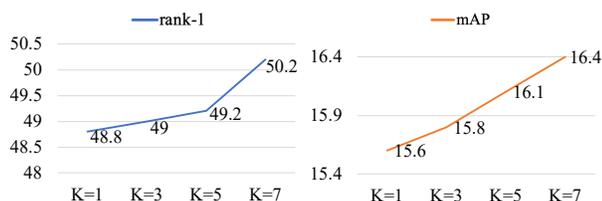


Figure 6. Performance on LSPS with different stripe number K .

stage lasts 120k, with learning rate as $1e-4$ and is decayed by 0.1 at 80k and 100k, respectively. RSFE is trained for 60k with learning rate as $1e-3$, decayed at 30k and 50k, respectively. BBA on three datasets are trained for 30k with learning rate as $1e-3$, and is decayed at 20k.

6.2. Ablation Study

Validity of Partial Matching. APNet addresses misalignment issue with partial matching based on stripe features. This part demonstrates the validity of this strategy. We first modify the Market-1501 [32] by adding random occlusions to query and gallery images. Based on this modified dataset, we compare different feature matching strategies and summarize the results in Table 3, where “global matching” does not differentiate occlusions in feature extraction and “partial matching” extracts features from valid regions.

Table 3 shows that, occlusion is harmful for reID, *e.g.*, degrades the mAP of global feature from 53.5% to 13.6%. Extracting stripe features from the entire bounding box boosts the reID performance. This shows the validity of part feature. Because partial matching extracts features from visible regions, it achieves better performance than global matching. Among three features used in partial matching, features extracted from mutual stripes achieve the best performance. It outperforms the feature from mutual region by large margins, *e.g.*, 30.9% vs. 21.4% in mAP. It also significantly outperforms the global feature in global matching. The above experiments hence shows the validity of the idea illustrated in Fig. 1, where features on the mutual visible regions are matched for similarity computation.

Features	PRW		LSPS	
	mAP	rank-1	mAP	rank-1
global	34.2	75.8	14.4	47.7
stripe	39.1	79.1	13.0	41.6
VPM [21]	40.0	80.2	16.0	49.5
stripe (BBA)	40.8	81.0	16.4	50.2
RSFE (BBA)	41.9	81.4	17.1	51.6

Table 4. Validity of BBA and RSFE in stripe feature extraction. Global refers to global feature extracted from detected bounding box. “stripe (BBA)” denotes extracting stripe features with vanilla part feature extractor based on BBA output. “RSFE (BBA)” denotes stripe feature extracted by RSFE.

Validity of BBA: BBA performs bounding box refinement and valid part estimation. We proceed to verify the validity of BBA in person search. Table 4 shows the performance of global feature and stripe feature before applying BBA. It is clear that, extracting valid stripe features from BBA refined bounding boxes, substantially boosts the reID performance. For instance, “stripe (BBA)” achieves mAP of 40.8% on PRW, better than the 34.2% and 39.1% of original global feature and stripe feature, respectively. We also show performance of the reproduced VPM [21], which is a recent method for partial matching. As shown in Table 4, stripe feature extracted with BBA performs better than VPM on both PRW and LSPS. This experiment hence shows the validity of BBA in bounding box refinement and valid part extraction.

Validity of RSFE: RSFE is designed to alleviate the negative effects of noisy regions illustrated in Fig. 4 (b). It is also designed to refine each part feature. Table 4 compares the stripe feature extracted by RSFE, *i.e.*, “RSFE (BBA)” against the one extracted by vanilla part feature extractor, *i.e.*, “stripe (BBA)”. It can be observed that, RSFE is important in boosting the stripe feature performance. For example, RSFE boosts the mAP by 1.1% and 0.7% on PRW and LSPS, respectively. The combination of BBA and RSFE achieves the best performance and outperforms the VPM by large margins. More comparisons with person search works will be presented in the next section.

Effects of part number K : We proceed to study the influence of part number K , and summarize the results in Fig. 6. $K = 1$ degrades the partial feature to global feature. The maximum horizontal stripe number K equals to the height of T , *i.e.*, 7 in our experiment. The figure shows that, stripes with finer scale help to boost the performance. This might due to the fact that larger K improves the feature resolution and robustness to occlusions and noises.

Discussions: To show the effects of BBA in bounding boxes refinement, we visualize bounding boxes before and after refinement in Fig. 7. The results show that, BBA effectively shifts the original bounding boxes to cover holistic body regions. This procedure eliminates misalignment errors and guarantees aligned stripe feature extraction.



Figure 7. Visualization of refined bounding boxes by BBA. Red and green boxes denote bounding boxes before and after refinement. BBA effectively estimates the holistic body region to eliminate misalignment errors.

Methods	# params (M)	speed (sec.)	TFLOPs
OIM	36	0.254	0.383
APNet	67	0.256	0.397

Table 5. Comparison of parameter and computational complexity between APNet and baseline OIM [24]. Speed and TFLOPs are measured on the NVIDIA 2080Ti GPU.

Methods	Reference	CUHK-SYSU		PRW	
		mAP	rank-1	mAP	rank-1
OIM [24]	CVPR2017	75.5	78.7	21.3	49.9
NPSM [4]	ICCV2017	77.9	81.2	24.2	53.1
CLSA [28]	ECCV2018	87.2	88.5	38.7	65.0
MGTS [1]	ECCV2018	83.0	83.7	32.6	72.1
CGPS [29]	CVPR2019	84.1	86.5	33.4	73.6
QEEPS [15]	CVPR2019	88.9	89.1	37.1	76.7
RDLR [3]	ICCV2019	93.0	94.2	42.9	72.1
APNet		88.9	89.3	41.9	81.4

Table 6. Comparison with recent works on CUHK-SYSU and PRW, respectively.

Besides the detector branch based on baseline OIM, APNet introduces additional BBA and RSFE modules. We compare the parameter and computational complexity between OIM and APNet in Table 5. The comparison shows that APNet achieves promising performance with comparable speed with baseline OIM, *e.g.*, 0.397 TFLOPs of APNet vs. 0.383 TFLOPs of OIM. Although the BBA and RSFE modules bring more parameters to APNet, they do not degrade its computational speed substantially. APNet is likely to be faster than person search works that treat detection and reID in separated steps, and is faster than some works like QEEPS [15], which compares each query-gallery pair for person search.

6.3. Comparison with recent works

CUHK-SYSU. We experiment on CUHK-SYSU with gallery size 100. APNet achieves the rank-1 accuracy of 89.3% and mAP of 88.9%, outperforming most of recent works. Note that RDLR [3] uses a stronger backbone ResNet50-FPN as well as ranking-based loss. Methods like CLSA [28], MGTS [1], and RDLR solve detection and reID with two separate models, which are expensive in both computation and storage. Compared with those works, our AP-

LSPS	OIM [24]	VPM [21]	APNet	APNet+v
mAP	14.4	16.0	17.1	18.8
rank-1	47.7	49.5	51.6	55.7

Table 7. Comparisons with recent works on LSPS. APNet denotes our method. APNet+v considers extra vertical stripe features.

Net is a unified model and is likely to present better computational and memory efficiencies.

PRW. On PRW, APNet achieves competitive performance, *e.g.*, 81.4% in rank-1 accuracy and 41.9% in mAP, outperforming most of recent works. APNet also significantly outperforms the RDLR [3] in rank-1 accuracy with a weaker backbone, *i.e.*, by 9.3% in rank-1 accuracy. Since certain query images in PRW cover partial body parts, APNet exhibits more substantial advantages with partial matching. It also outperforms CGPS [29] and QEEPS [15] by 7.8% and 4.7% in rank-1, respectively. Note that, CGPS [29] and QEEPS [15] input each query-gallery pair into CNN for similarity computation, hence also present inferior retrieval efficiency compared with APNet.

LSPS. We finally present experimental results on LSPS. OIM [24] is compared as the baseline. APNet brings 3.9% rank-1 improvement over the baseline, achieving 51.6% and 17.1% in rank-1 and mAP, respectively. APNet also outperforms the recent VPM [21], which solves the partial reID problem. To consider the misalignment in horizontal direction, we further apply additional vertical stripe features. The corresponding method, *i.e.*, APNet+v achieves the best performance, achieving 55.7% and 18.8% in rank-1 and mAP, respectively. It is also clear that, the performance of APNet on LSPS is substantially lower than the ones on PRW and CUHK-SYSU, indicating the challenges of LSPS.

7. Conclusion

This paper proposes an APNet to solve the bounding box misalignment issues in person search. APNet refines detected boxes to cover the estimated holistic body regions, meanwhile extracting part features from visible body parts. This formulates reID as a partial feature matching procedure, where valid part features are selected for similarity computation, while part features on occluded or noisy regions are discarded. This paper also contributes a LSPS dataset, which is by far the largest and the most challenging dataset for person search. Experiments show that APNet brings considerable performance improvement on LSPS. Moreover, it achieves competitive performance on existing person search benchmarks like CUHK-SYSU and PRW.

Acknowledgments This work is supported in part by The National Key Research and Development Program of China under Grant No. 2018YFE0118400, in part by Beijing Natural Science Foundation under Grant No. JQ18012, in part by Natural Science Foundation of China under Grant No. 61936011, 61425025, 61620106009, 61572050, 91538111.

References

- [1] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *ECCV*, 2018.
- [2] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.
- [3] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-id driven localization refinement for person search. In *ICCV*, 2019.
- [4] Liu Hao, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Zhao Bo, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. In *ICCV*, 2017.
- [5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [7] Lingxiao He, Liang Jian, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *CVPR*, 2018.
- [8] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.
- [9] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, 2019.
- [10] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018.
- [11] Jianing Li, Shiliang Zhang, Qi Tian, Meng Wang, and Wen Gao. Pose-guided representation learning for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [12] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.
- [13] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *ECCV*, 2016.
- [14] Jiayu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, 2019.
- [15] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. In *CVPR*, 2019.
- [16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.
- [17] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.
- [18] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry S Davis, and Wen Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *CVPR*, 2015.
- [19] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, 2016.
- [20] Chi Su, Shiliang Zhang, Fan Yang, Guangxiao Zhang, Qi Tian, Wen Gao, and Larry S Davis. Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping. *Pattern Recognition*, 66:4–15, 2017.
- [21] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *CVPR*, 2019.
- [22] 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 *ECCV*, 2018.
- [23] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, 2019.
- [24] Xiao Tong, Li Shuang, Bochao Wang, Lin Liang, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017.
- [25] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*, 2018.
- [26] Guanshuo Wang, Yufeng Yuan, Chen Xiong, Jiwei Li, and Zhou Xi. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 2018.
- [27] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*, 2017.
- [28] Lan Xu, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In *ECCV*, 2018.
- [29] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, and Xiaokang Yang. Learning context graph for person search. In *CVPR*, 2019.
- [30] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *CVPR*, 2019.
- [31] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *CVPR*, 2019.
- [32] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.
- [33] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, and Qi Tian. Person re-identification in the wild. In *CVPR*, 2017.
- [34] Meng Zheng, Srikrishna Karanam, Ziyang Wu, and Richard J Radke. Re-identification with consistent attentive siamese networks. In *CVPR*, 2019.
- [35] WeiShi Zheng, Li Xiang, Xiang Tao, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *CVPR*, 2016.
- [36] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13, 2018.