Weakly Supervised Person Search

Lan Yan
State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences
Beijing, China
School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
yanlan2017@ia.ac.cn

Wenbo Zheng
School of Software Engineering, Xi'an Jiaotong University
Xi'an, China
State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences
Beijing, China
zwb2017@stu.xjtu.edu.cn

Fei-Yue Wang
State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences
Beijing, China
feiyue.wang@ia.ac.cn

Chao Gou
School of Intelligent Systems Engineering, Sun Yat-sen University
Guangzhou, China
gouchao@mail.sysu.edu.cn

***Abstract*—While existing person search methods have achieved good performance, they require the images used for training contain labels about the identity and bounding box location of each person. However, it is expensive and difficult to manually annotate these labels in the large scale scenario. To overcome this issue, we consider weakly supervised person search. The weakly supervised setting means during training we only know which identities appear in the image set and how many individuals present in each image, without any identity or location information on the image. Facing this challenge, we propose a clustering and patch based weakly supervised learning (CPBWSL) framework, which separately addresses two sub-tasks including pedestrian detection and person re-identification. Particularly, we introduce multiple detectors to provide more detection results as well as fuzzy c-means clustering algorithm to cluster these results and remove low membership ones. Moreover, a patch based learning network is designed to generate different patches and learn discriminative patch features. Extensive experiments on two benchmarks indicate that the proposed weakly supervised setting is feasible and our method can achieve performance comparable to some fully supervised person search methods.**

***Index Terms*—Person search; Pedestrian detection; Person re-identification; Weakly supervised learning**

I. Introduction

Person search aims to search for a target person in a gallery of whole scene images [1]. It has obtained increasing attention in recent years, owing to its tremendous possibilities for security and video surveillance applications. This problem is challenging because of the existence of various distractions in real scenarios, such as cluttered background, occlusion, changing illumination, camera viewpoint and low resolution.

This work is supported in part by the Key Research and Development Program of Guangzhou (202007050002) and the National Natural Science Foundation of China (61533019, 61806198, U1811463). *(Corresponding author: Chao Gou)*

Until now, some efforts have been devoted to address this problem [1]–[13]. However, these methods are presented for person search in a fully supervised setting. As shown in Fig. 1(a), the images used in fully supervised learning are carefully manually annotated, and each of them contains several bounding boxes and identification labels. Notwithstanding this fully supervised paradigm contributes to learn a robust person search model, its scalability and usability in the practical and large-scale scenarios are severely limited due to the high cost of the data-labeling process. To overcome the shortcoming, we consider designing a person search method which could work with weak supervision.

Hence, in this paper, we would like to study the person search problem in a weakly supervised setting. Under this setting, an annotator is only required to roughly look at the raw image set and determine the number of persons in each image and which identities appear in the set, but not to label any identity or bounding box on the images. In other words, only the labels regarding the presence of the identity and the number of individuals are provided, and yet more detail ground-truth indicating pixel-level bounding box and in which picture the identity arise is not available. Specifically, as is evident in Fig. 1(b), in the gallery set, the first image is annotated by "2 Persons" indicating that number of pedestrians is two in this image. The gallery set is tagged with a label "{Person 1, Person 2, Person 3, Person 4}" implying that Person 1, Person 2, Person 3 and Person 4 are present in this set. Obviously, these annotations are weak.

More specially, the weak supervision in our setting falls into the category of inexact supervision [14], where only coarse-grained labels are provided for training. Compared with the conventional supervised learning setting, in such a setting, the annotating costs for person search can be significantly

978-1-7281-8206-3/20/$31.00 ©2020 IEEE
DOI 10.1109/DSAA49011.2020.00031

(a) Fully supervised setting

(b) Weakly supervised setting

Fig. 1. Illustration of two settings. (a) Fully supervised setting: images in gallery set are manually labeled and each image has bounding boxes and corresponding identity labels. Note that the pale yellow bounding boxes in the figure represent unknown identities. (b) Weakly supervised setting: every picture in gallery set has a label describing the number of people in the picture, and the gallery set has a label suggesting which identities present in the set, while the annotations regarding pixel-level bounding box and which picture the identity arise in are not given.

mitigated as well as the scalability and availability of the person search model in real-world application scenarios can be improved. We term this setting as *weakly supervised person search*.

Given a query set of images with probe persons, our objective is searching the gallery and discovering the images where the probe individual appears. To this end, we propose a novel **c**lustering and **p**atch **b**ased **w**eakly **s**upervised **l**earning (CPBWSL) framework for weakly supervised person search. Conventionally, person search contains two sub-tasks, i.e., pedestrian detection and person re-identification. The former deems all people as one class and aims to distinguish people from the background while the latter regards different person as different categories and aims to identify these different categories. Therefore, sharing representations between these two sub-tasks is not appropriate and we solve them separately. Moreover, instead of using a single detector, we leverage multiple detectors to provide richer bounding boxes. Simultaneously, we introduce the fuzzy c-means (FCM) clustering algorithm [15] to remove the false detection results. Then, to learn discriminative feature for person re-identification, we design a patch based learning network that is able to generate different patches and learn discriminative patch features. We perform experiments on two popular benchmarks including CUHK-SYSU [1] and PRW [16]. The results indicate the feasibility of our weakly supervised setting and effectiveness of the proposed weakly supervised person search method.

To summarize, the main contributions of this work are twofolds:

- We propose a new and practical problem—weakly supervised person search. *To the best of our knowledge, this work is the first to present and investigate the weakly supervised problem in person search.*
- We develop a novel clustering and patch based weakly supervised learning (CPBWSL) framework to address the weakly supervised person search problem. It achieves the encouraging performances of 77.6% mAP and 33.2% mAP on CUHK-SYSU and PRW datasets respectively, which are comparable to some fully supervised person search methods.

II. Related Works

In this section, some studies about person re-identification are introduced firstly. Afterwards, we present the related works about person search.

A. Person Re-identification

Person re-identification is a feature matching problem essentially. Early solutions focus on designing hand-crafted features [17]–[21] and learning distance metrics [22]–[27]. For example, Zhao *et al.* [20] present using the combination of SIFT feature and color histogram feature for person re-identification. Kostinger *et al.* [22] propose the KISSME approach which addresses the person re-identification issue by matrix distances learning. These early solutions realize some success on small datasets. Nevertheless, the hand-crafted feature based methods usually fail in large-scale matching since the representation capabilities of manual features are still limited, and the traditional distance metrics learning methods tends to over-fit the training data. Hence, the performance of these early methods is unsatisfactory.

In recent years, due to the significant advantage of deep learning in retrieval accuracy, person re-identification methods based on deep learning have attracted extensive attention. Most of them can be divided into two categories. The first category [28]–[33] focuses on building robust and discriminative image representations. Li *et al.* [31] propose to extract the most discriminative features by using the attention mechanism. Sun *et al.* [33] employ the idea of self-supervision and present a visibility-aware part model to learn the region-level features. The second category [34], [35] aims to learn a deep metric for person re-identification. Build upon the triplet loss, Chen *et al.* [34] present a quadruplet loss to acquire a more effective deep metric.

However, the person re-identification methods introduced above are all supervised, and they require precise annotations whose labeling process is expensive. Moreover, considering that the performance of unsupervised approaches is usually limited by the absence of explicit supervised information, Meng *et al.* [36] introduce the weakly supervised person re-identification problem which only have video-level labels and design a cross-view multi-instance multi-label learning approach to solve it.

It is worth noting that person re-identification aims to extract and match features based on given bounding boxes. In a

practical application, the person re-identification approaches should be combined with off-line detection methods.

B. Person Search

Person search has attracted extensive research interests since the two large-scale databases CUHK-SYSU and PRW were published. In addition to collecting the CUHK-SYSU dataset, Xiao *et al.* [1] first propose to jointly handle pedestrian detection and person re-identification with a single end-to-end network which is trained with Online Instance Matching (OIM) loss. Liu *et al.* [2] recursively search and refine the location of the target person on the panoramic images. Chang *et al.* [4] present a deep reinforcement learning based method without proposal computing to solve the person search problem. Xiao *et al.* [7] propose a individual aggregation network (IAN) and introduce a center loss to increase the feature discriminative power. Yan *et al.* [8] consider the underlying relationship between persons in the scene image and introduces a relative attention module to adaptively search and filter informative context the scene. Munjal *et al.* [37] jointly optimize the detection and re-identification parts and present query-guidance for OIM, which is provided by performing person search on the uncropped query image.

Unlike OIM, Lan *et al.* [5] perform pedestrian detection and person re-identification separately. they consider the multi-scale matching problem in person search and proposes a Cross-Level Semantic Alignment (CLSA) deep learning approach to tackle it. Through the systematic comparison between the separation models which separately conduct detection and re-identification and the joint models which perform them jointly, Cheng *et al.* [6] find that the separation approaches can improve the performances of detection and re-identification. Accordingly, they select the separation scheme and design a mask-guided two-stream CNN model (MGTS) for person search which has one stream to model foreground person and the other stream for processing the original image. Considering the characteristics of joint approaches and separation approaches, Han *et al.* [9] propose to learn the detector and re-identification model in an end-to-end manner without sharing features. Li *et al.* [38] focus on studying the time bottleneck of person search and present a real-time pipeline.

Note that while some of these aforementioned approaches has achieved great performance, all of them study person search in a fully supervised scenario, relying on costly annotations. To overcome this issue, our work is devoted to solving the weakly supervised person search problem.

III. The Proposed Method

In this section, we first provide an overview of the proposed weakly supervised person search framework. Then, we introduce more details for each part in our framework individually. Finally, we show how our approach works in the inference procedure the training is complete.

A. Overview

Our goal is learning an effective and robust model in the weakly supervised setting to find a target person in the gallery set of whole scene images, given a query set of images with probe individuals. As illustrated in Fig. 2, the proposed framework separately deals with the pedestrian detection and person re-identification. Given a whole scene image input, we first fed it into M detectors and acquire a number of bounding boxes. Then, we set the number of clusters as the number of people in the panoramic image (in Fig. 2 the specific value is three), and use the FCM clustering algorithm to cluster these bounding boxes. After that, if a bounding box has a low degree of membership to any class, the bounding box will be discarded. Subsequently, at the person re-identification stage, under the guidance of two loss functions, the patch based learning network which mainly consists of the feature extractor and the patch generation network is used to learn discriminative features for person re-identification. Next, we will detail how to implement the aforementioned procedures.

B. Detection and Clustering

Pedestrian detection is important for accurate person search [1], [16]. Since under the proposed weakly supervised person search setting, it is impractical to learn a great detection model from scratch, instead of designing a pedestrian detector specially, we apply the existing excellent detection model as our pedestrian detector. However, due to the existence of uncontrolled false alarms and miss-detections, detection is still a problem that has not been completely solved, the results of a single detection model would not provide sufficient support to subsequently learn a accurate person re-identification model. Hence, we consider using multiple detectors to enrich detection results for better person search.

Although the multiple detectors provide much more detection results, they also give more incorrect bounding boxes. Since, in the weakly supervised scenario, we have no annotation to tell our detectors which bounding box is incorrect and how to exactly distinguish people from backgrounds and other objects with similar appearances, a intuitionistic idea is employing a clustering method.

Considering that we have no way of knowing the exact value of the wrong or correct bounding boxes, that is, they are fuzzy. To this end, we use a classic fuzzy clustering approach, i.e., the FCM clustering algorithm, to obtain the degree of membership of each bounding box sample to each cluster, and remove the samples with low membership. More specifically, the idea can be formulated by minimizing an objective function below:

$$J_q = \sum_{n=1}^{N}\sum_{f=1}^{F}(u_{nf})^q \left\|\mathbf{b}_n - \mathbf{c}_f\right\|^2, 1 \le q < \infty \qquad (1)$$

where N is the number of samples, F is the number of clusters, $\mathbf{b}_n$ is the n-th sample, $\mathbf{c}_f$ denotes the centroid of the f-th cluster and u_{nf} denotes the membership of the n-th sample to the f-th cluster. q is a weighting exponent and it controls the relative weights placed on each of the squared errors. For most cases, setting q to 2 or 3 is a good choice [15]. Hence, in our experiment we set $q = 2$.

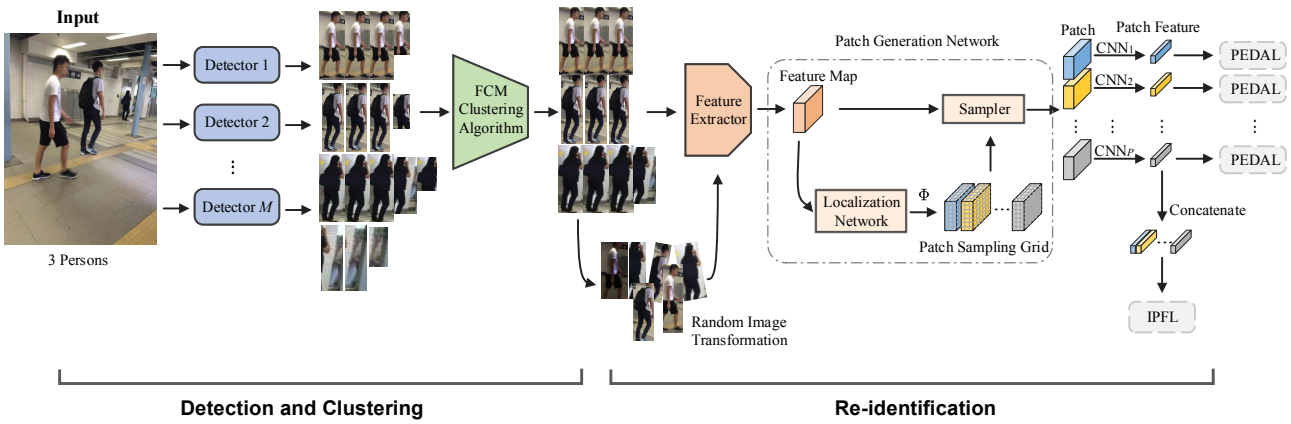

Fig. 2. The proposed framework. First, a whole scene image is inputed into M detectors and some detected boundary boxes are obtained. Then, we use the FCM clustering algorithm to cluster these bounding boxes and remove low membership ones. The cropped images used in person re-identification are acquired. Finally, with the guidance from PEDAL and IPFL, the patch based learning network which mainly consists of the feature extractor and the patch generation network is leveraged to learn discriminative features for person re-identification.

The objective function can be minimized by updating and iterating the following two expressions:

$$\mathbf{c}_f = \frac{\sum_{n=1}^{N}(u_{nf})^q \mathbf{b}_n}{\sum_{n=1}^{N}(u_{nf})^q} \tag{2}$$

$$u_{nf} = \frac{1}{\sum_{z=1}^{F}\left(\frac{\|\mathbf{b}_n - \mathbf{c}_f\|}{\|\mathbf{b}_n - \mathbf{c}_z\|}\right)^{\frac{2}{q-1}}} \tag{3}$$

The iteration stops when the membership matrix U satisfies $\left\|U^{(k+1)} - U^{(k)}\right\| < \varepsilon$. Here, k is the number of iterations, ε is iteration termination parameter, in our experiment we set $\varepsilon = 1 \times 10^{-6}$.

Particularly, in our framework, the number of samples N is the number of bounding boxes for all detected people on a panoramic image. The number of clusters F is equal to the number of people appearing in the image, which is provided in the weakly supervised person search setting. $\mathbf{b}_n$ is actually the coordinate of n-th bounding box sample.

After the iteration, we obtain the final membership matrix U. Then, we find the maximum membership degree of bounding box sample $\mathbf{b}_n$ in the matrix U and denote it by u_{nf^*}. If the maximum membership degree u_{nf^*} is lower than a given threshold, we will remove the bounding box $\mathbf{b}_n$. We cluster the detection results of each image respectively and cut off low-membership detections by a given threshold. Only the remaining bounding boxes are applied by the re-identification network.

C. Patch Based Person Re-identification

After detection results are acquired, we aim to extract discriminative features for person re-identification. Although we do a number of processing in detection stage to get more accurate bounding boxes, there is still a large gap compared with the elaborate manual labels. Since some researchers have proposed to study local discriminative feature [30], [39] for person re-identification and achieve better performances than those global feature learning models [40], [41], we consider extracting discriminative local features and reducing the reliance on exact bounding boxes.

Specially, without any identity information, we design a patch based learning network for person re-identification, which mainly consists of a feature extractor and the patch generation network. Considering that feature maps are more effective and generally smaller than the image and benefit the reduction of network computation and complexity [30], [31], we use a feature extractor to get the feature maps from the cropped images. Similar with spatial transformation networks [42], the patch generation network which can produce abundant patches from the feature map, is composed of a localization network, a sampler and patch sampling grids, as shown in Fig. 2. The localization network consists of a convolutional layer and two fully connected layers. Given a feature map input, it predicts P spatial locations parameterized by a group of affine transformation arguments $\Phi = [\phi_1, ..., \phi_p, ..., \phi_P]$. With a predict transformation parameter ϕ_p, a patch sample grid can be computed. Such a grid contains the coordinates of sampling points to sample patches from the input feature map. Subsequently, using the sampler, we can obtain P patches for each image. For more details about the patch generation network, please refer to [42].

It is noteworthy that for each input feature map the patch generation network generates P patches located in different spatial regions which may involve different parts of the body with diverse semantics [30], [39]. Therefore, we use P different CNN branches to encode these P different patches, and each branch independently learns the discriminative features.

As we aim to distinguish different persons, given the unlabeled features of patch, it is natural to achieve this goal by enhancing inter-class discrepancy and the intra-class compactness simultaneously. To this end, we introduce the

patch-based discriminative feature learning loss (PEDAL) [43] to push the dissimilar patches away and meanwhile pull similar patches together. We denote the p-th patch feature of i-th cropped image by $\mathbf{x}_i^p$. To find similar patches, comparing $\mathbf{x}_i^p$ to all patches of the other cropped images is necessary. Considering that it is intractable to perform in the optimization of deep learning [1], we store the patch features and maintain a memory bank W^p, where $W^p = \{\mathbf{w}_l^p\}_{l=1}^L$ and L is the number of the cropped images. During training, we update $\mathbf{w}_l^p$ by:

$$\mathbf{w}_{l,t}^p = \begin{cases} (1-r) \times \mathbf{w}_{l,t-1}^p + r \times \mathbf{x}_{l,t}^p, & t > 0, \\ \mathbf{x}_{l,t}^p, & t = 0, \end{cases} \tag{4}$$

where t denotes the training epoch, and $r \in [0, 1]$ denotes the update rate, $\mathbf{x}_{l,t}^p$ denotes the latest patch feature. When $t = 0$, the memory bank is initialized firstly. When $t > 0$, it updating as the training goes on, and $\mathbf{w}_{l,t}^p$ can be regarded as the online approximation of $\mathbf{x}_i^p$ [1].

Then, given $\{\mathbf{w}_l^p\}_{l=1}^L$, we can compute the l_2 distance between $\mathbf{x}_i^p$ and each $\mathbf{w}_l^p$ and get k nearest patches of $\mathbf{x}_i^p$, whose set is denoted by $\mathcal{K}_i^p$. The PEDAL can be expressed as:

$$\mathcal{L}_{\text{PEDAL}}^p = -\log \frac{\sum_{\mathbf{w}_l^p \in \mathcal{K}_i^p} e^{-\frac{s}{2}\left\|\mathbf{x}_i^p - \mathbf{w}_l^p\right\|_2^2}}{\sum_{l=1, l \neq i}^L e^{-\frac{s}{2}\left\|\mathbf{x}_i^p - \mathbf{w}_l^p\right\|_2^2}} \tag{5}$$

where s denotes the scaling factor. By minimizing $\mathcal{L}_{\text{PEDAL}}^p$, the model is encouraged to pull similar patches $\mathcal{K}_i^p$ closer to $\mathbf{x}_i^p$ and push dissimilar patches $\{\mathbf{w}_l^p | \mathbf{w}_l^p \notin \mathcal{K}_i^p\}$ away form $\mathbf{x}_i^p$ in the feature space. Accordingly, the patch features learned by the model would be more discriminative.

Additionally, to further mine the potential image-level identification information, we introduce the patch feature learning loss (IPFL) [43], which leverages all the patch features of a image to offer image-level guidance. As is evident in Fig. 2, we concatenate the patch features of the same cropped image and form a new feature denoted as $\mathbf{x}_i$. We randomly transform the cropped image to generate a proxy positive sample. These random image transformations include cropping, rotation along with contrast, saturation and brightness changing. Simultaneously, we use cyclic ranking [43] to mine the hardest negative sample feature $\mathbf{n}_i$ for $\mathbf{x}_i$. The IPFL can be formulated by:

$$\mathcal{L}_{\text{IPFL}} = \max\{\|\mathbf{x}_i - \mathbf{p}_i\|_2 - \|\mathbf{x}_i - \mathbf{n}_i\|_2 + m, 0\} \tag{6}$$

where m is a margin of this loss and $\mathbf{p}_i$ denotes a proxy positive sample feature.

Therefore, the total loss function for person re-identification can be defined as:

$$\mathcal{L} = \mathcal{L}_{\text{IPFL}} + \lambda \frac{1}{P} \sum_{P}^{p=1} \mathcal{L}_{\text{PEDAL}}^p \tag{7}$$

where λ controls the impact of the PEDAL.

D. Inference Procedure

At test phase, our aim is to find a target person in the gallery set of whole scene images, given a query set of images with probe individuals. Just as our model is composed of pedestrian detection and person re-identification, the inference process is also divided into two stages.

The inference procedure of our proposed method is shown in Fig. 3. Taking the CUHK-SYSU testing dataset as an example, under our weakly supervised setting, the gallery set contains 6978 images and a label "{2900 IDs}" suggesting 2900 identities present in the set, every image in the gallery set has a label describing the number of people in the image. For instance, in Fig. 3, the image in the lower left corner has a tag "{12 Persons}" which denotes there are 12 persons appearing in this image.

As illustrated in Fig. 3, in stage one, pedestrian detection and FCM clustering are conducted on each scene image. After clustering, we get filtered bounding boxes for each gallery image. Then, in stage two, we feed all filtered bounding boxes for the whole gallery set as well as the probe image into the patch based person re-identification network. According to the patch-based apparent features extracted from inputs, the re-identification network can find the bounding boxes most similar to probe image, i.e., the matched bounding boxes. Note that when we get the bounding boxes in the stage one, the corresponding scene image information of each bounding box like name is also saved, so once we get the matching bounding boxes, the final results are naturally obtained.

IV. Experiments and Results

In this section, we first introduce two commonly used person search benchmarks and the evaluation metrics. Afterwards, we show our implementation details and conduct comparison experiments. Finally, we perform a number of ablation experiments to study the impacts of introducing multiple detectors and the FCM clustering algorithm.

A. Datasets and Evaluation Protocol

To verify the effectiveness of our approach, we conduct experiments on both CUHK-SYSU [1] and PRW [16] benchmarks. To quantitatively evaluate our approach, we select the top-1 matching rate metric and the mean Average Precision (mAP) as performance measurements, which are widely adopted for evaluating person re-identification and person search methods. The top-1 matching rate metric regards person search as a ranking and matching problem. Only when the overlap rate between the boundary box in the top-1 prediction box and the ground-truth is greater than the threshold 0.5, it is counted as a match. The mAP reflects the accuracy of detecting the query person from all gallery images.

CUHK-SYSU. CUHK-SYSU is a large-scale person search dataset with a wide variety of scenes which consists of 18184 scene images and 96143 annotated pedestrian bounding boxes. As reported in Table I, the dataset has 8432 labeled identities and the rest annotated pedestrians are unknown identities. Moreover, it provides the official train/test split, i.e., a training

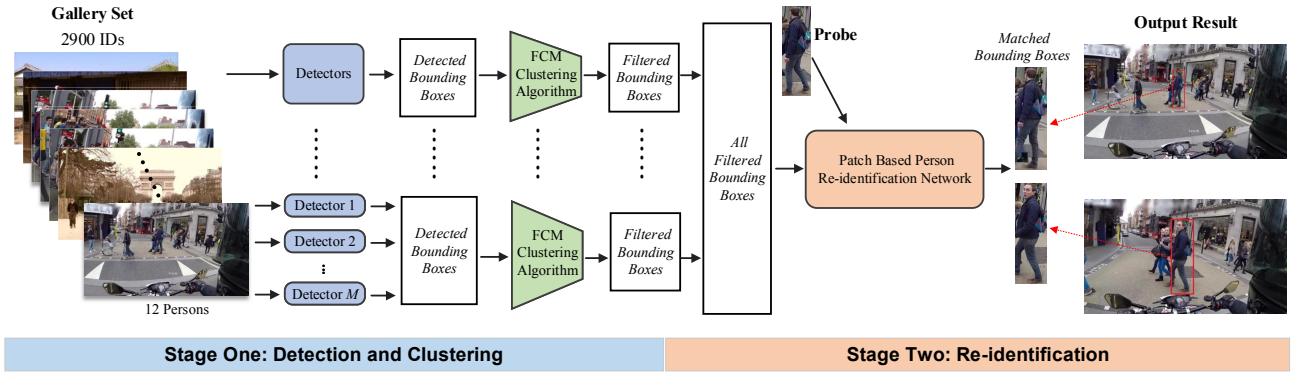

Fig. 3. The inference procedure. There are two stage: in stage one, pedestrian detection and FCM clustering are conducted on each scene image, and the filtered bounding boxes for each gallery image are obtained; in stage two, all filtered bounding boxes for the whole gallery set as well as the probe image are fed into the patch based person re-identification network. Finally, the re-identification network outputs the matched bounding boxes and the corresponding scene images.

TABLE I
DATA STATISTICS AND EVALUATED SETTING OF THE CHUK-SYSU AND PRW DATABASES. BBOX: BOUNDING BOX.

Datasets	*Overall*			*Training*			*Testing*		
	Images	Bboxes	IDs	Images	Bboxes	IDs	Images	Bboxes	IDs
CHUK-SYSU	18184	96143	8432	11206	55272	5532	6978	40871	2900
PRW	11816	43110	932	5704	18048	482	6112	25062	450

set with 11206 images and 5532 identities as well as a testing set with 6978 gallery images and 2900 probe persons.

PRW. PRW is captured from six cameras on a university campus. As shown in Table I, it contains 11816 frames and 43110 pedestrian bounding boxes. Among these pedestrians, 932 identities are labeled and the rest are regarded to the unknown persons. The dataset is officially split into a training set with 5704 frames and 482 identities and a testing set that includes a gallery of 6112 frames and 2057 probe persons with 450 identities.

B. Implementation Details

We implement our framework with Pytorch [44] and conduct all experiments on two NVIDIA TITAN XP GPUs. We choose three off-the-shelf detectors from the open source object detection toolbox mmdetection [45], including Faster-RCNN [46], RetinaNet [47] and Cascade R-CNN [48]. They are pre-trained on the MS COCO dataset [49] and use ResNet-101 [50] with FPN as their backbone network. 3. All detectors are initialized by the pretrained model and then frozen. The membership threshold is empirically set to 0.8.

For the feature extractor, we use ImageNet-pretrained ResNet-50 [50] and remove the fully connected layer along with set the stride of last residual unit as 1. We pre-train the batch based learning network with the MSMT17 dataset [51] and fix the patch generation network during training on other datasets. As for the loss functions, we follow the setting of [43] and the update rate r in Eq.(4) is set to 0.1, the scaling factor s is set to 5, the margin m and the weight λ are set to 2. During the person re-identification model training process, the cropped images are resized to 384×128. We use SGD [52] optimizer for training, with a batch size of 40 and the momentum of 0.9. The learning rate is initialized at 1×10^{-4} and decayed by 0.1 every 40 epochs. We train the re-identification model on the unlabeled cropped images for 50 epochs.

C. Performance Comparison

In this subsection, we report the performance evaluation results on both CUHK-SYSU and PRW datasets. As far as we know, there is no other weakly supervised person search method, so we compare our method with the existing fully supervised person search methods, including OIM [1], NPSM [2], RCAA [4], MGTS [6], CLSA [5], IAN [7], GCNPS [8] and QEEPS [37]. Except the above approaches, we also choose some other approaches for comparision which combine different pedestrian detectors (DPM [53], ACF [54], LDCF [55], CCF [56], and R-CNN [57]) and person descriptors (BoW [58], LOMO [23], DSIFT [20]) and distance metric (KISSME [22], XQDA [23]).

Comparison on CUHK-SYSU. Table II demonstrates the comparative results on CUHK-SYSU with a gallery size of 100. Following the notations defined in [1], we employ "CNN" to represent Faster R-CNN [46] based on ResNet-50 and "IDNet" to denote the re-identification module in OIM.

As can be seen from the Table II, well-designed person search approaches such as OIM and IAN are better than all the combination of traditional approaches. Our method achieves 77.6% mAP and 88.7% in top-1 matching rate metric, and it outperforms OIM and IAN with ResNet-50 backbone by 2.1% and 1.3% in mAP, 10.0% and 8.6% in top-1 matching rate,

TABLE II

COMPARISION WITH THE FULLY SUPERVISED PERSON SEARCH METHODS ON THE CUHK-SYSU DATASET WITH A GALLERY SIZE OF 100. IAN (R-50) DENOTES IAN WITH RESNET-50 BACKBONE, AND IAN (R-101) DENOTES IAN WITH RESNET-101 BACKBONE. THE "FULLY" TYPE MEANS THE METHOD IS FULLY SUPERVISED. SIMILARLY, "WEAKLY" INDICATES THAT IT IS A WEAKLY SUPERVISED PERSON SEARCH METHOD.

Method	Type	mAP(%)	top-1(%)
ACF [54] + DSIFT [20] + Euclidean	Fully	21.7	25.8
ACF [54] + DSIFT [20] + KISSME [22]		32.3	38.1
ACF [54] + LOMO [23] + XQDA [23]		55.5	63.1
CCF [56] + DSIFT [20] + Euclidean	Fully	11.3	11.7
CCF [56] + DSIFT [20] + KISSME [22]		13.4	13.9
CCF [56] + LOMO [23] + XQDA [23]		41.2	46.4
CCF [56] + IDNet [1]		50.9	57.1
CNN [46] + DSIFT [20] + Euclidean	Fully	34.5	39.4
CNN [46] + DSIFT [20] + KISSME [22]		47.8	53.6
CNN [46] + BoW [58] + Cosine		56.9	62.3
CNN [46] + LOMO [23] + XQDA [23]		68.9	74.1
CNN [46] + IDNet [1]		68.6	74.8
OIM [1]	Fully	75.5	78.7
IAN (R-50) [7]		76.3	80.1
IAN (R-101) [7]		77.2	80.5
NPSM [2]		77.9	81.2
RCAA [4]		79.3	81.3
MGTS [6]		83.0	83.7
GCNPS [8]		84.1	86.5
QEEPS [37]		84.4	84.4
CLSA [5]		**87.2**	88.5
Ours	Weakly	77.6(7^{th})	**88.7**(1^{st})

TABLE III

COMPARISION WITH SEVERAL FULLY SUPERVISED PERSON SEARCH METHODS ON THE PRW DATASET. THE TYPE "FULLY" REPRESENTS THAT THE METHOD IS FULLY SUPERVISED. "WEAKLY" MEANS THAT IT IS WEAKLY SUPERVISED.

Method	Type	mAP(%)	top-1(%)
ACF-Alex [54] + LOMO [23] + XQDA [23]	Fully	10.3	30.6
ACF-Alex [54] + IDE$_{det}$ [16]		17.5	43.6
ACF-Alex [54] + IDE$_{det}$ [16] + CWS [16]		17.8	45.2
DPM-Alex [53] + LOMO [23] + XQDA [23]	Fully	13.0	34.1
DPM-Alex [53] + IDE$_{det}$ [16]		20.3	47.4
DPM-Alex [53] + IDE$_{det}$ [16] + CWS [16]		20.5	48.3
LDCF [55] + LOMO [23] + XQDA [23]	Fully	11.0	31.1
LDCF [55] + IDE$_{det}$ [16]		18.3	44.6
LDCF [55] + IDE$_{det}$ [16] + CWS [16]		18.3	45.5
OIM [1]	Fully	21.3	49.9
IAN (R-50) [7]		23.0	61.9
NPSM [2]		24.2	53.1
MGTS [6]		32.6	**72.1**
Ours	Weakly	**33.2**(1^{st})	60.1(3^{rd})

respectively. In addition, our approach has better performance than the state-of-the-art CLSA in top-1 matching rate. These imply that our weakly supervised setting is viable and the proposed method can yield competitive performance compared with the fully supervised person search approaches.

Comparison on PRW. Furthermore, we compare our approach with 13 fully supervised competitors on the PRW dataset. Nine of them are combinations of detectors and re-identification methods discussed in [16]. Comparison results are listed in Table III, where "DPM-Alex" represents DPM [53] + AlexNet [59]-based R-CNN, "IDE$_{det}$" denotes ID-discriminative Embedding, and "CWS" is the Confidence Weighted Similarity defined in [16].

Compared to the CUHK-SYSU dataset, the PRW dataset contains fewer images and identities but has more bounding boxes per identity (36.8 in PRW versus 2.8 in CUHK-SYSU), which makes it more challenging. Hence, as shown in Table III, we can observe that all the specially designed approaches have worse performance on the PRW dataset than on the CUHK-SYSU dataset, particularly the mAP. However, on the PRW dataset, our method also achieve competitive performances on both mAP and top-1 matching rate metics. It is noteworthy that while the mAP of our method is lower than that of MGTS on the CUHK-SYSU dataset, on the more

TABLE IV
THE RESULTS OF OUR ABLATION EXPERIMENT ON THE CUHK-SYSU DATASET WITH GALLERY SIZE OF 100. THE METHODS ABOVE THE DASHED LINE USE THE SINGLE DETECTOR.

Detection schemes	mAP(%)	top-1(%)
Cascade R-CNN	34.2	44.2
Faster R-CNN	37.6	47.8
Retinanet	42.3	51.1
Multiple detectors	68.1	80.9
Multiple detectors + FCM (Ours)	**77.6**	**88.7**

challenging PRW dataset, the our model gains the promotion of 0.6% on mAP compared with MGTS. It is shown that our model is more robust and scalable than MGTS. This consistently indicates the feasibility of the proposed weakly supervised setting along with the effectiveness and proves the scalability of our model.

D. Ablation Study

In this subsection, we conduct several ablation experiments on the CUHK-SYSU dataset to validate the effectiveness of introducing the multiple detectors and the FCM clustering algorithm. The results are reported in Table IV. The methods above the dotted line only leverage a single existing detector and do not use the FCM clustering algorithm. "Multiple detectors" represents a variant of our model, which uses multiple detectors to get the bounding boxes without FCM clustering of these bounding boxes. It is noteworthy that other settings remain unaltered during the ablation study.

The effectiveness of multiple detectors. From Table IV, we can observe that simply using a single detector can not achieve satisfactory results. The person search method with multiple detectors outperforms the methods which only use a single detector by a large margin, in both mAP and top-1 matching rate metrics. This demonstrates that employing the multiple detectors for providing richer bounding boxes is helpful to improve the performance of our model. Therefore, introducing the multiple detectors is effective for the task of weakly supervised person search.

V. CONCLUSIONS

In this work, we first propose to study person search in the weakly supervised scenario. Our motivation is that the fully supervised person search methods have a huge demand for fine manual annotation, which is extremely time-consuming and laborious, and leads to poor scalability of these algorithm. In a weakly supervised setting, only labels indicating which identities appear in the image set and how many persons present in each image are required, without any identity or location information on the image. We develop a novel clustering and patch based weakly supervised learning (CPBWSL) framework to tackle the weakly supervised person search problem, which copes pedestrian detection and person re-identification separately. During the pedestrian detection stage, multiple detectors and the FCM clustering algorithm are introduced to offer better detection results for person search. While in person re-identification stage, we design a patch based learning network to learn the discriminative patch features. Experimental results validate the feasibility of our weakly supervised setting for person search and show the superiority of our model.

REFERENCES

[1] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR))*, 2017, pp. 3415–3424.

[2] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan, "Neural person search machines," in *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 2017, pp. 493–501.

[3] Z. He and L. Zhang, "End-to-end detection and re-identification integrated net for person search," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 349–364.

[4] X. Chang, P.-Y. Huang, Y.-D. Shen, X. Liang, Y. Yang, and A. G. Hauptmann, "Rcaa: Relational context-aware agents for person search," in *European conference on computer vision(ECCV)*, 2018, pp. 84–100.

[5] X. Lan, X. Zhu, and S. Gong, "Person search by multi-scale matching," in *European conference on computer vision(ECCV)*, 2018, pp. 536–552.

[6] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream cnn model," in *European conference on computer vision(ECCV)*, 2018, pp. 734–750.

[7] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, and J. Feng, "Ian: the individual aggregation network for person search," *Pattern Recognition*, vol. 87, pp. 332–340, 2019.

[8] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2019, pp. 2158–2167.

[9] C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang, C. Gao, and N. Sang, "Re-id driven localization refinement for person search," in *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 2019, pp. 9814–9823.

[10] W. Dong, Z. Zhang, C. Song, and T. Tan, "Bi-directional interaction network for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, June 2020.

[11] ——, "Instance guided proposal network for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, June 2020.

[12] D. Chen, S. Zhang, J. Yang, and B. Schiele, "Norm-aware embedding for efficient person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, June 2020.

[13] C. Wang, B. Ma, H. Chang, S. Shan, and X. Chen, "Tcts: A task-consistent two-stage framework for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, June 2020.

[14] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2017.

[15] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[16] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017, pp. 1367–1376.

[17] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 2007, pp. 1–8.

[18] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision(ECCV)*. Springer, 2008, pp. 262–275.

[19] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2360–2367.

[20] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2013, pp. 3586–3593.

[21] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision(ECCV)*. Springer, 2016, pp. 499–515.

[22] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2012, pp. 2288–2295.

[23] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2015, pp. 2197–2206.

[24] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 2015, pp. 3765–3773.

[25] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 2015, pp. 3685–3693.

[26] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 356–370, 2016.

[27] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2015, pp. 1846–1855.

[28] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017, pp. 3960–3969.

[29] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proceedings of the IEEE international conference on computer vision(ICCV)*, 2017, pp. 3219–3228.

[30] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *European conference on computer vision(ECCV)*, 2018, pp. 480–496.

[31] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018, pp. 2285–2294.

[32] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018, pp. 5363–5372.

[33] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR))*, 2019, pp. 393–402.

[34] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017, pp. 403–412.

[35] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[36] J. Meng, S. Wu, and W.-S. Zheng, "Weakly supervised person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2019, pp. 760–769.

[37] B. Munjal, S. Amin, F. Tombari, and F. Galasso, "Query-guided end-to-end person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2019, pp. 811–820.

[38] J. Li, F. Liang, Y. Li, and W.-S. Zheng, "Fast person search pipeline," in *ICME*. IEEE, 2019, pp. 1114–1119.

[39] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *European conference on computer vision(ECCV)*, 2018, pp. 402–419.

[40] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 2017, pp. 3800–3808.

[41] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018, pp. 4320–4328.

[42] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[43] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2019, pp. 3633–3642.

[44] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[45] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[46] 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*, 2015, pp. 91–99.

[47] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 2017, pp. 2980–2988.

[48] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018, pp. 6154–6162.

[49] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision(ECCV)*. Springer, 2014, pp. 740–755.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2016, pp. 770–778.

[51] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018, pp. 79–88.

[52] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *ICML*, 2013, pp. 1139–1147.

[53] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[54] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[55] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Advances in neural information processing systems*, 2014, pp. 424–432.

[56] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 2015, pp. 82–90.

[57] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2014, pp. 580–587.

[58] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 2015, pp. 1116–1124.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.