Bottom-Up Foreground-Aware Feature Fusion for Person Search

Wenjie Yang
wenjie.yang@nlpr.ia.ac.cn
CRISE, CASIA & School of Artificial Intelligence, UCAS
Beijing, China

Dangwei Li
dangwei.li@nlpr.ia.ac.cn
CRISE, CASIA & School of Artificial Intelligence, UCAS
Beijing, China

Xiaotang Chen
xt.chen@nlpr.ia.ac.cn
CRISE, CASIA & School of Artificial Intelligence, UCAS
Beijing, China

Kaiqi Huang*
kaiqi.huang@nlpr.ia.ac.cn
CRISE, CASIA & School of Artificial Intelligence, UCAS
Beijing, China

ABSTRACT

The key to efficient person search is jointly localizing pedestrians and learning discriminative representation for person re-identification (re-ID). Some recently developed *task-joint* models are built with separate detection and re-ID branches on top of shared region feature extraction networks, where the large receptive field of neurons leads to background information redundancy for the following re-ID task. Our diagnostic analysis indicates the *task-joint* model suffers from considerable performance drop when the background is replaced or removed. In this work, we propose a subnet to fuse the bounding box features that pooled from multiple ConvNet stages in a bottom-up manner, termed bottom-up fusion (BUF) network. With a few parameters introduced, BUF leverages the multi-level features with different sizes of receptive fields to mitigate the background-bias problem. Moreover, the newly introduced segmentation head generates a foreground probability map as guidance for the network to focus on the foreground regions. The resulting foreground attention module (FAM) enhances the foreground features. Extensive experiments on PRW and CUHK-SYSU validate the effectiveness of the proposals. Our **B**ottom-**U**p **F**oreground-Aware Feature **F**usion (BUFF) network achieves considerable gains over the state-of-the-arts on PRW and competitive performance on CUHK-SYSU.

CCS CONCEPTS

• **Computing methodologies** → **Biometrics**; *Image representations*; Object detection.

KEYWORDS

Person Search; Pedestrian Detection; Person Re-Identification; Feature Learning; Attention

ACM Reference Format:
Wenjie Yang, Dangwei Li, Xiaotang Chen, and Kaiqi Huang. 2020. Bottom-Up Foreground-Aware Feature Fusion for Person Search. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413991

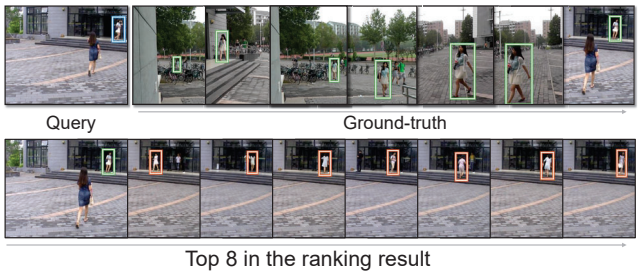


Figure 1: Illustration of the background-bias problem. The first row shows the scene images with the bounding boxes of a person in PRW, where blue and green denote query and ground truth. The second row shows the *Top-8* output matches given by the baseline, in which identification head and detection head share the RoI features. Bounding boxes with green and red boundary denote true positive and false positive, respectively. As we can observe, the false positives are in the same scene as the query, which indicates that the re-ID features encode redundant background information.

1 INTRODUCTION

Person search aims at retrieving a certain person among video frames, which involves both pedestrian localization and re-ID. As a subtask, re-ID [30, 39] is a technique that matches person images across cameras. It has wide application in video surveillance, such as person retrieval [30, 39], multi-target-multi-camera tracking [26, 37] and activity recognition [21], *etc.* Although re-ID has achieved significant progress in recent years [4, 10, 28, 33], the gap between this task and real application is still considerable. One reason is that current re-ID methods are carried out upon high-quality bounding boxes either from filtered detections or human annotations [18, 31, 38]. In real-world systems, detection results are far from perfect, which include false alarms, missing detection and inaccurate localization. To eliminate the gap, researchers [34, 40] propose to tackle detection and re-ID in a joint framework, also known as the task of person search.

*Also with CAS Center for Excellence in Brain Science and Intelligence Technology.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413991

Main-stream approaches to person search can be categorized into two groups of *task-joint* and *task-cascaded* methods, differentiated by learning re-ID representation from the feature maps of scene images or from the cropped image of pedestrian boxes. The *task-joint* methods [2, 20, 23, 34, 36] develop a multi-task model, in which detection and re-ID share the backbone. Upon the backbone, there are two branches. The pedestrian detection branch performs bounding boxes binary classification and regression, which are then used to pool Region of Interest (RoI) features from the backbone (*a.k.a* RoI Pooling [9] in detection) and send it to a separate re-ID branch for identification features extraction. The *task-cascaded* methods [3, 12, 16, 40] formulate it as a cascaded framework. In the pipeline, a pedestrian detector first takes a full-size video frame as input and produces bounding boxes of pedestrians. Then these image patches are cropped from the frame, resized, and fed to a followed re-ID model for extracting identification features. Although the joint framework sets the need for coping with task compatibility, bounding box scales, as well as optimization balance, the *task-joint* group enjoy the advantage of saving model parameters and fast inference [13].

Our work falls into the *task-joint* group and shares the backbone between detection and re-ID tasks. In this paper, we identify a factor caused by previous implementation that has negative impact on re-ID feature learning. As illustrated in Fig. 1, we analyze failure cases during testing and discover that when searching in the gallery set, The *task-joint* baseline model tends to return those bounding boxes with scene background similar to or exactly the same as the query one, although the identities are different from, and dissimilar to, the query person. To investigate the influence of the background information to existing *task-joint* methods, we create background-influence datasets [29] using the foreground masks which generated by setting the regions inside the bounding boxes to 1 while others to 0. We found that the commonly used *task-joint* baseline suffers from considerable performance drop when the background is removed or replaced. This indicates that the re-ID features encoding excessive information of the scene context, which is unexpectedly overwhelming and conceals the identity information within the bounding box. The features of RoI that sent to the re-ID head are pooled from the backbone, which has a large receptive field and has the ability to capture information beyond the bounding box. Consequently, in the feature space, the same background pull closer pedestrians of different identities while distinct background push away those of the same identity. It could be the reason that causes the problem illustrated in Fig. 1. For disentangled feature learning and robustness to background variation, we argue that the final re-ID features of a bounding box should not capture the relevance between background and the box.

Since the receptive field of the neurons becomes larger as the ConvNets[6, 17] deepen, it is intuitive to exploit the features of shallower layers for mitigating the background-bias. In this paper, the goal of the proposed bottom-up fusion subnet is to leverage the ConvNet's multi-scale features while pursuing semantic alignment at the same scale. To achieve this goal, we rely on an architecture that fuses large receptive field, semantically strong features with small receptive field, semantically weak features via a bottom-up pathway and transform layers (Fig. 3(c)). Moreover, with the weak box-wise annotated masks for training, the newly introduced segmentation head generates foreground probability map as guidance to help the network focusing on the regions of foreground. Extensive experiments on the original and background-influence datasets demonstrate the effectiveness of our proposed methods on mitigating the background-bias.

The main contributions of this paper can be summarized as threefold. 1) A bottom-up fusion subnet is proposed to fuse features with different size of receptive field from multiple ConvNet stages, so as to mitigate the background-bias problem. 2) A foreground attention module is proposed to guide the network to learn discriminative representation by focusing on the visual appearances of pedestrians. 3) Extensive experimental results show that the proposals achieve the new state-of-the-art performance on PRW and competitive performance on CUHK-SYSU.

2 RELATED WORKS

In this section, we review literatures on pedestrian detection and person search.

Pedestrian Detection Traditional pedestrian detectors generally involve hand-crafted features and linear classifiers. Representative methods include HOG [5], DPM [8] and ACF [7], *etc.* Along with the development of deep learning, extensive CNN based models have been developed to achieve fast and accurate detection. The pioneering work R-CNN [9] adopts selective search to generate proposals, CNN to extract features for each, and SVM to perform classification. The following work Fast R-CNN [9] improves detection efficiency largely by sharing feature maps of proposals. To make the proposal generation tunable by the loss function, an end-to-end framework Faster R-CNN [25] is devised. It is a widely adopted model in many applications, including person search [3, 16, 20, 23, 34, 36]. The detector in our framework is also based on Faster R-CNN.

Person Search The re-ID benchmarks [18, 31, 38] are based on hand-cropped person images or filtered detections, which makes it far from practical application. In real scenario, a system is required to search a person among video frames, where detection errors could be detrimental to re-ID algorithms. Person search is the task to meet this requirement of a jointly optimized framework and two large-scale benchmarks PRW [40] and CUHK-SYSU [34] are introduced. Xu *et al.* [35] first introduces this task with sliding window searching based on hand-crafted features. Zheng *et al.* [40] analyze various combinations of detectors and re-ID methods, as well as how detection could be beneficial to assist final ranking. **Task-joint Methods** To reach an efficient and jointly optimized approach for person search, Xiao *et al.* [34] extend an identification head parallel to bounding box classifier of Faster R-CNN. Liu *et al.* [20] develop a neural person search machine that starts from the full gallery image and iteratively narrow its region of interest till the person is correctly located. Chang *et al.* [2] utilize a similar searching framework, where a reinforcement learning based agent is trained to perform location. To ensure robust person search under large gallery size, Yan *et al.* [36] emphasize neighboring co-travelers, with a context graph to model the global similarity of probe-gallery pairs. Munjal *et al.* [23] utilize query image as guidance in a thorough way when searching the target person in a gallery image. They implement query context modulated backbone, query-relevant proposals

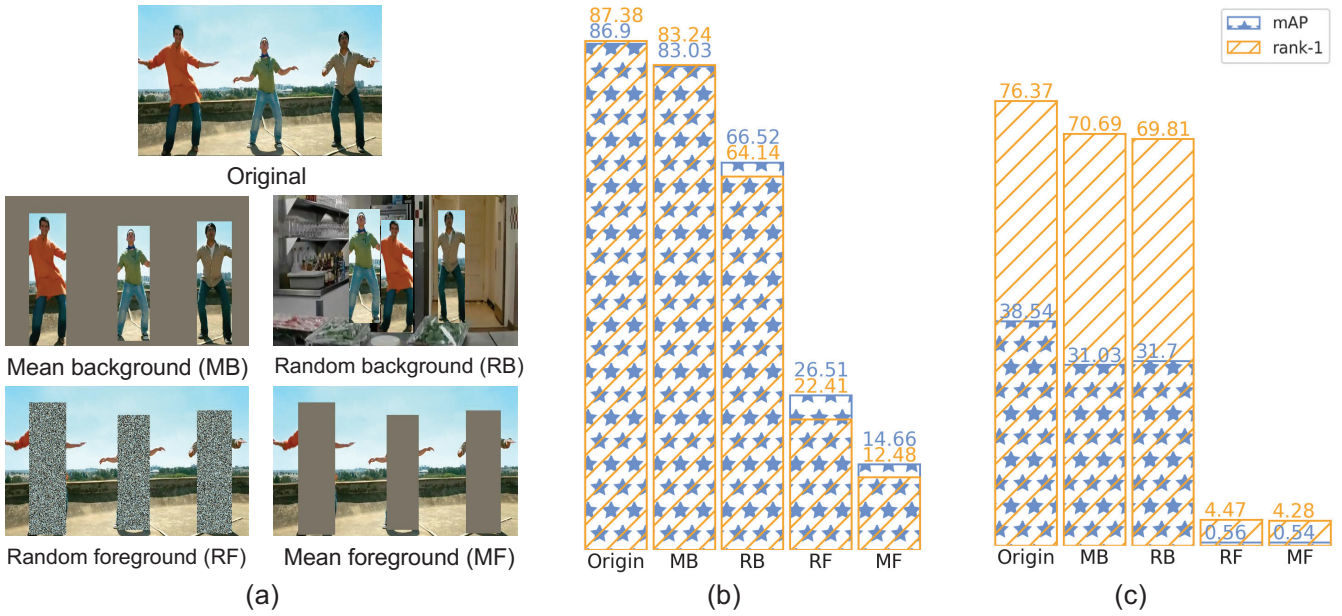


Figure 2: (a) Examples of the original CUHK-SYSU and the synthesized background-influence datasets. (b) mAP and rank-1 results on CUHK-SYSU. The baseline model is trained on the original CUHK-SYSU, then evaluated on the five datasets. (c) Analogous results on PRW. We use the ground-truth bounding boxes for evaluation.

and re-ID similarity score based non-maximum suppression (NMS).

Task-cascaded Methods Lan *et al.* [16] consider the wide range of scales in detected persons and construct an in-network feature pyramid to solve the multi-scale matching problem. Chen *et al.* [3] develop a two-stream feature extraction network operating on both original bounding box and foreground of the box to obtain enriched representation. These methods use separate pre-trained pedestrian detectors and only learn the re-ID networks, thus are not end-to-end trainable. Recently, Han*et al.* [12] propose to refine localization using re-ID training loss, in which way person-to-person clutter can be reduced while accessory could be appropriately encompassed for re-ID feature extraction.

3 INVESTIGATIONS ON BACKGROUND-BIAS

In this section, we conducts quantitative analysis on the background-bias problem. We firstly create four types of datasets based on the existing dataset (3.1), then the commonly used baseline model is evaluated on the synthesized datasets to investigate the influence of background on re-ID (3.2).

3.1 The background-influence datasets

The mean background (MB), random background (RB), random foreground (RF), and mean foreground (MF) datasets are created based on the original dataset and the bounding boxes annotations. Examples of the datasets are illustrated in Fig. 2 (a).

Let the regions outside the boxes be background, and the regions inside the boxes be foreground. **MB** images are obtained by setting the background region to the mean pixel values of ImageNet [6], while keeping the foreground (pedestrians) unchanged. On the contrary, **MF** image sets foreground to the mean pixel values but keeps the background region. The visual appearances of the pedestrians are removed. **RF** image fills the foreground region with the randomly sampled pixels from the original image. As for the **RB** dataset, we adopt different strategies for PRW and CUHK-SYSU. On PRW, the images are collected from six immobile cameras, it is thus convenient to obtain the six *pure-background* images in which there are no pedestrians. We paste the bounding boxes into a randomly sampled *pure-background* image captured from a different camera. On CUHK-SYSU, most of the images are captured from a moving camera, the background varies a lot but the same identity usually lies in the same background. We simply paste the probe bounding box into the location of a bounding box in a different scene image, and keep the gallery set unchanged.

3.2 Influence of background to re-ID

We conduct experiments based on the baseline model (4.1) and use the ground-truth bounding boxes for test. The accuracies of the model trained on the original dataset are shown in Fig. 2 (b) and (c). On *MB* dataset, with only the background removed, the rank-1 accuracy decreases by 4.1% and 5.7% on CUHK-SYSU and PRW, respectively. On *RB* dataset, the background has changed, both the mAP and rank-1 performance drop significantly. For example, the rank-1 accuracy drop by a large margin of nearly 23% from 87.38% to 64.14% on CUHK-SYSU (from *Origin* to *RB* in Fig. 2 (b)). On *RF* dataset, the foreground is full of random noise that humans cannot distinguish, but the model achieves high mAP accuracy of 22.41% on CUHK-SYSU. On *MF* dataset, although the discriminative visual appearances of the pedestrians are removed, the rank-1 accuracies

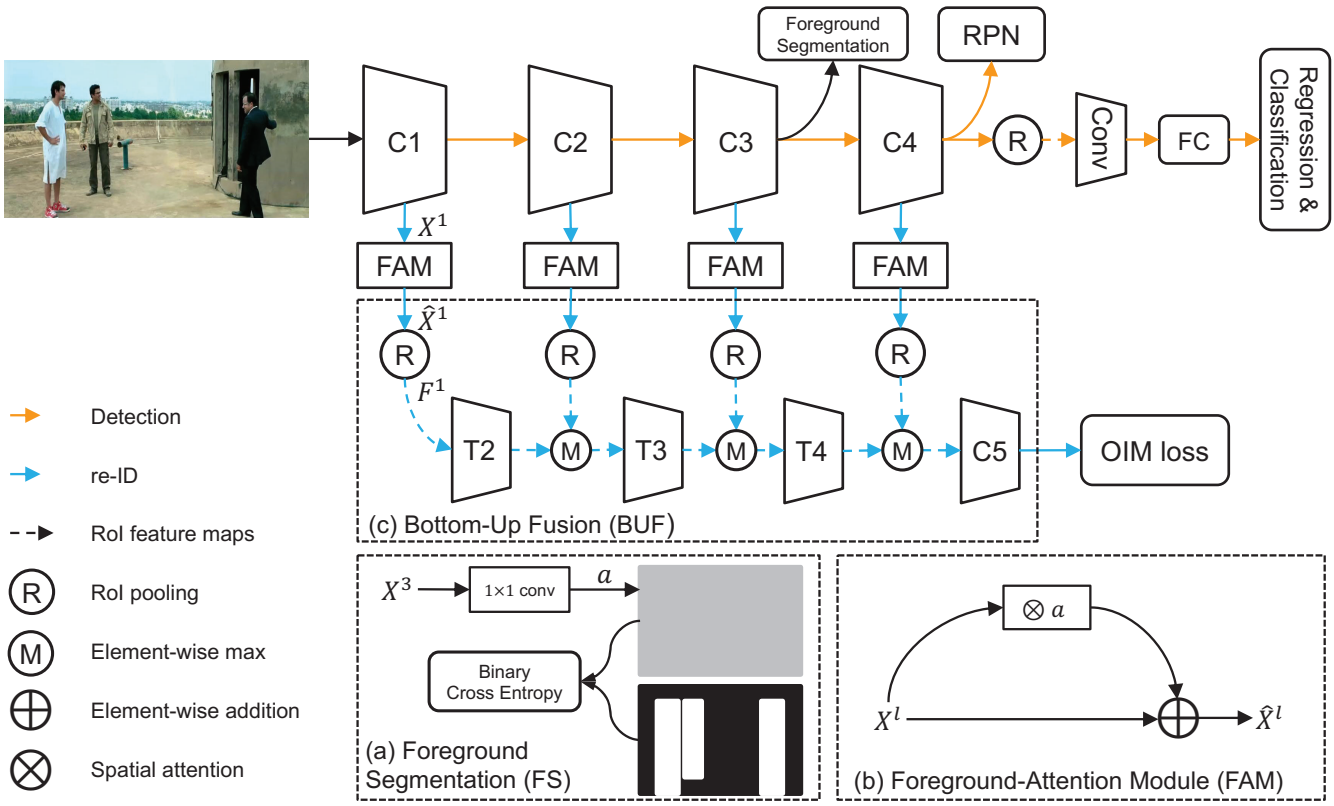


Figure 3: The proposed BUFF is a multi-task network for joint pedestrian detection (red dataflow) and re-ID presentation learning (blue dataflow). The input scene image forwards through the *conv1* to *conv4* of ResNet50 and produces feature maps X^l at l-th layer. In (a), X^3 is converted by a 1×1 *conv* layer to obtain the foreground probability map a. Then in (b), the FAM takes X^l and a as input and produce $\hat{X}^l$ in a residual manner. In the bottom-up fusion stream, i.e., (c), the RoI pooling layer is connected to $\hat{X}^l$ for pooling the RoI features F^l. After that, F^{l-1} is transformed by *T*-l to have the same shape with F^l, then the RoI features are merged by element-wise max. The transform layer *T*-l is a residual block that consists of three convolutional layers and a residual connection. The RPN on the top of *conv4* generates proposals for the all the RoI pooling layers.

on CUHK-SYSU and PRW much higher than $\frac{1}{\#bboxes}$, where *#bboxes* is the number of ground-truth bounding boxes in gallery.

The large receptive field of RoI features of higher semantic level biases the re-ID feature towards capturing redundant relevance between background and the visual appearances of pedestrians. It makes the trained model overfitting on the original dataset, but fail in real-world application where the background of the same peron can be quite different. Therefore, it is necessary to improve the re-ID feature extraction mechanism in the person search framework.

4 THE PROPOSED FRAMEWORK

This section presents the technical details on the proposed model termed BUFF. As shown in Fig. 3, BUFF builds two streams for detection and re-ID, which aims to pursue these two tasks in a joint optimization framework (Sec. 4.1). With the need of only weak box-wise annotated masks for training, the foreground segmentation head generates foreground probability map as guidance to help the network focusing on the regions of foreground (Sec. 4.2). Finally, a novel bottom-up fusion network is proposed to exploit the features of the shallow layers, so as to mitigate the background-bias and learn discriminative re-ID representation (Sec. 4.3).

4.1 Baseline

We first introduce the baseline model used in this paper. The baseline is built with a shared Region Feature Extraction (RFE) network followed by two separate heads, i.e., a detection head and a re-ID head. As illuminated in the detection stream of Fig. 3, the RFE network consists of *conv1* to *conv4* of ResNet50 [14], and a standard Region Proposal Network (RPN) [25]. Given an input image, the *conv4* layer produces a tensor $T \in \mathbb{R}^{d \times h \times w}$, which can be interpreted as dense d channels of feature maps with $h \times w$ spatial resolutions. The RPN is bulit on T to detect pedestrian candidates, where a $512 \times 3 \times 3$ convolutional layer is added to transform T and 9 anchors is assigned at each spatial location of the transformed feature maps. Then a softmax classifier and a linear layer are respectively performed for bounding boxes binary classification and regression. After Non-Maximum Suppression (NMS), the RPN outputs 128 proposals for the following RoI pooling layer, which

produces RoIs feature maps of spatial resolutions 14×7, denoted as $F \in R^{128\times d\times 14\times 7}$.

In the detection head, i.e., RCNN, F passes through a *conv* layer and produces $F_{det} \in R^{128\times 256\times 7\times 4}$. F_{det} is firstly reshaped to $\hat{F}_{det} \in R^{128\times \hat{d}}$, where $\hat{d} = 256 \times 7 \times 4$, then a fully-connected (FC) layer is applied to extract detection features $f_{det} \in R^{128\times 256}$. Finally, a softmax classifier is used to distinguish between person and non-person, and a linear regression is applied to refine the bounding boxes locations.

In the re-ID head, the *conv5* of ResNet50 is connected to F and outputs high sematic level feature, which is denoted by $F_h \in R^{128\times 2048\times 7\times 3}$. Then forwards F_h through global average pooling (GAP) layer to produce feature $f_h \in R^{128\times 2048}$ as re-ID representations. In inference, the L_2-normalized re-ID feature are used for computing cosine similarities.

4.2 Foreground-Aware Feature Learning

We propose foreground-aware feature learning to reduce the relevance between background and the visual appearances of pedestrians. The foreground segmentation head produces a foreground probability map indicating the likelihood of residing on pedestrian or background. Then we approach foreground attention module (FAM) which uses the foreground probability map to *illuminates* the foreground in the input feature maps for the following re-ID feature learning.

Foreground Segmentation. We add foreground segmentation head to *conv3* in the detection stream. The segmentation head consists of a single convolutional layer with 1×1 kernel and a Sigmoid activation layer. It takes as input feature maps of *conv3*, i.e., X^3, and outputs a single channel foreground probability map, denoted as a. We use the abundance of bounding box annotations available in person search datasets to generate weak segmentation ground truth masks. Given a scene image and the corresponding bounding boxes annotations, it is convenient to obtain the box-wise masks by set the regions inside boxes to 1 while others to 0. As this work [1] points out, pixel-wise mask not offer a significant advantage over box-wise when placing the segmentation head at a higher level of the network. Since the ground-truth mask has been pooled significantly, the differences between these two kinds of annotations is diminished. The loss function that used to train the segmentation head is the binary cross-entropy, given by

$$L_{seg} = -\frac{1}{N}\sum_{i=1}^{N} y_i * log(x_i) + (1 - y_i) * log(1 - x_i) \tag{1}$$

where y_i is the annotation label, x_i is the foreground probability and N denotes the number of spatial resolutions.

Foreground Attention Module (FAM). Given the foreground probability map a generated by the foreground segmentation head, it is feasible to use it as the guidance for foreground attention. As illustrated in the bottom right of Fig. 3, i.e., (c), we put the attention mechanism inside the residual branch in a residual block. Specifically, given the feature maps X^l of *conv-l* and foreground probability map a, FAM outputs $\hat{X}^l$ as the foreground-aware re-ID features. Formally, $\hat{X}^l$ is computed as

$$\hat{X}^l = X^l + X^l \otimes a^l \tag{2}$$

where $\otimes$ denotes spatial attention. This mechanism help the network focusing on pedestrians in the input images for learning more discriminative features for describing person visual appearance.

4.3 Bottom-Up Fusion (BUF)

Since the receptive field of the neurons becomes larger as the network deepens, it is intuitive to exploit the features of the shallower layers for learning background-insensitive representation. The goal of BUF is to leverage the ConvNet′s multi-scale features, which have the receptive field from small to large, and build the fusion architecture.

The feed-forward of the detection backbone computes a feature hierarchy, which consists of feature maps at several scales with a scaling step of 2. Before the BUF, the multi-scale feature maps pass through FAM to form the foreground-aware features, denoted as $\{\hat{X}^1, \hat{X}^2, \hat{X}^3, \hat{X}^4\}$. The construction of our fusion architecture involves a bottom-up pathway and feature transforms, as introduced in the following. For each positive proposal generated by the RPN, the corresponding RoI feature patches $F^l \in R^{d\times h\times w}$ is extracted from $\hat{X}^l$ using RoI pooling. After that, a transform layer is connected to F^{l-1} and outputs $\hat{F}^{l-1}$ which has the same shape with F^l. The transform layer is a residual block consists of three convolutional layers and a residual connection. Then $\hat{F}^{l-1}$ and F^l are merged by element-wise max. We call the transform layer and the followed element-wise max a *fusion stage*. Each *fusion stage* merges two RoI feature maps that of different semantic levels. The feature map from a shallower layer is of low-level semantic, and its discriminability is enhanced via the transform layer. Finally, the aggregated feature F^a, i.e., the output of *conv5*, forwards through global average pooling layer to produce $f \in R^{2048}$ as final re-ID representation. With only a few parameters (about 2M) introduced, the BUF achieves significant improvement on both of the two datasets.

4.4 Objective Function

The identification loss used in this paper is the Online Instance Matching (OIM) [34] loss, which is a commonly used loss function in person search. The pedestrian detection model consists of RPN and the detection head, i.e., the *conv* and FC layer. Following the standard training setting of object detection[25], the learning objective of detection can be written as follows,

$$L_{det} = L_{cls}^{rpn} + L_{reg}^{rpn} + L_{cls} + L_{reg} \tag{3}$$

where L_{cls} and L_{reg} are computed at the top of the detection head. The overall optimization objective for the learning scheme is given as follows,

$$L = L_{det} + L_{seg} + \lambda * L_{oim} \tag{4}$$

where λ is the trade-off hyperparameter and we use $\lambda = 0.5$ in all experiments.

5 EXPERIMENTS

The experiments are performed on two large-scale person search datasets, i.e., PRW[40] and CUHK-SYSU[34]. In this section, we first introduce the datasets, evaluation protocols, and some implementation details. Then ablation study is conducted to verify effectiveness

Table 1: Ablation study on the key components of BUFF. We report the mAP, rank-1 accuracies of re-ID and recall of the pedestrian detector. *Detected* indicates testing with the proposals produced by the detector, while *Labeled* means testing with the ground-truth bounding boxes. The gallery size of 100 and 4000 are reported on CUHK-SYSU, while the whole gallery set of PRW serves as the search space.

Method	PRW gallery size=6112			CUHK-SYSU gallery size=100			CUHK-SYSU gallery size=4000	
	mAP	rank-1	recall	mAP	rank-1	recall	mAP	rank-1
				Detected				
Baseline	36.3	74.5	96.6	84.3	84.8	98.9	65.4	69.0
+*BUF*	41.9	80.2	96.4	89.8	90.7	98.7	75.5	78.3
+*FAM*(BUFF)	42.2	81.0	96.7	90.7	91.6	98.7	77.6	80.1
				Labeled				
Baseline	38.5	76.4	100	86.9	87.4	100	68.3	71.3
+*BUF*	43.3	81.4	100	91.2	91.7	100	76.9	79.3
+*FAM*(BUFF)	44.4	82.4	100	92.2	92.8	100	78.9	81.6

of the components in our approach, followed by experimental results with comparison to the state-of-the-art methods.

Table 2: Training, gallery and query splits of PRW and CUHK-SYSU. #ped. w/ID denotes the number of pedestrian bounding boxes that annotated with IDs.

Dataset	split	#frame	#ID	#ped. w/ID	#ped. w/o ID
PRW	train	5,704	482	14,907	3,141
	gallery	6,112	450	19,127	5,935
	query	2,057	450	2,057	0
CUHK-SYSU	train	11,206	5,532	15,085	40,187
	gallery	6,978	2,900	8,345	32,526
	query	2,900	2,900	2,900	0

5.1 Datasets and Evaluation Protocols

The datasets statistics and split protocols are shown in Table 2. **PRW** contains 43,110 pedestrian bboxes in 11,816 scene images, which are captured by six cameras with different viewpoints in a university. **CUHK-SYSU** contains 18,184 scene images, labeled with 8,432 identities and 96,143 bounding boxes. The images either captured in an urban city by hand-held cameras or collected from movie snapshots. In the training set, CUHK-SYSU contains much more IDs and ID unknown bounding boxes (40,187 against 3,141 in PRW), but less bounding boxes per ID (2.7 against 31 in PRW). In test, the whole gallery set of PRW serves as the search space for each query, while gallery size of 100 is commonly used for CUHK-SYSU. **Evaluation protocol** For re-ID, we adopt the Cumulative Matching Characteristic (CMC) [11] and mean Average Precision (mAP) [38] as performance metrics. For pedestrian detection, Recall is used to measure the performance of pedestrian detector.

5.2 Implementation Details

Model. We adopt ResNet50 [14] that pre-trained on ImageNet [6] as the base model. It is noted that we follow the setting in OIM [34] that fix the first 7×7 convolution layer and the batch normalization (BN) layers as constant affine transformations, while keep the other BN layers as normal. In the re-ID stream, we adopt the first residual block of *conv2*, *conv3*, and *conv4* as *T2, T3*, and *T4*,respectively. The parameters are not shared with the detector. The weights of all the fully-connected layer are randomly initialized.

Data Preprocessing. The input scene images are re-scaled such that their shorter side is 600 pixels. For data augmentation, standard horizontal flipping is used during the training stage.

Optimization. We utilize PyTorch[24] to implement the overall framework. The stochastic gradient descent (SGD) optimizer with momentum of 0.9 is employed. The batch size and weight decay is set to 4 and 1×10^{-4} respectively. The number of proposals of each image is set to 64 and the fraction of positive proposal is 0.5. We adopt the anchor scales of {8, 16, 32} and aspect ratios of {1, 2}. The whole network is end-to-end trained for 5 epochs with a base learning rate of 0.001 which is reduced by a factor of 10 after the fourth epoch. On PRW (totally 11,408 training images, including the flipped images), the person search model consumes about 3.5 hours with a NVIDIA TITAN Xp GPU.

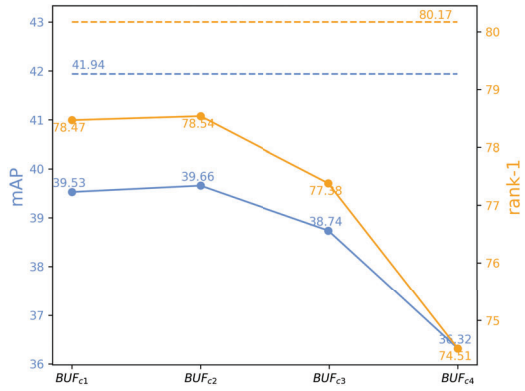
5.3 Ablation Study

In this section, we investigate the effectiveness of each component in the BUFF network by conducting analytic experiments on PRW and CUHK-SYSU. As shown in Table 1, we observe from the second row that BUF brings significant mAP and rank-1 promotion. For examples, the rank-1 accuracy is respectively increased by 5.7% (from 74.5% to 80.2%) and 6%(from 84.8% to 90.7%) on PRW and CUHK-SYSU. Such promotion benefits from two aspects. On the one hand, BUF combines the shallow layer′s features that capture less background information to form a more discriminative representation. On the other hand, the re-ID features are extracted from the newly introduced bottom-up fusion stream and re-ID and detection no longer share the features. Because the detection task treats the persons with distinct identities as the same category and learns the common representation of persons, while re-ID dedicates to distinguish these identities by learning fine-grained and discriminative

Table 3: The mAP and CMC rank-1 accuracies on the original and the background-influence datasets (Sec. 3.1). Note that the ground-truth bounding boxes are used for evaluation. *Bsl* is the baseline model.

Method	Origin		MB		RB	
	mAP	rank-1	mAP	rank-1	mAP	rank-1
	PRW					
Bsl	38.5	76.4	31.0	70.7	31.7	69.8
BUFF	44.4	82.4	39.0	79.3	39.1	77.4
	CUHK-SYSU					
Bsl	86.9	87.4	83.0	83.2	66.5	64.1
BUFF	92.2	92.8	87.2	87.9	81.3	82.0

Figure 4: Performances of BUF (dashed line) and BUF using single-level feature (solid line) on PRW. BUF_{c1} indicates BUF only takes the feature produced by *conv1* as input in both training and test stages.



features. We believe the efficiency and simplicity of our method will benefit future research of person search and related applications. The results of *+FAM* shows that the FAM achieves considerable accuracy margin of 0.3% and 0.9% in mAP, 0.9% and 0.9% in rank-1 on PRW and CUHK-SYSU, respectively. This demonstrates the importance of removing the background information outside the bounding boxes in the procedure of re-ID feature learning.

5.4 Results on Background-Influence Datasets

We further evaluated our BUFF on the mean-background (MB) and random-background (RB) datasets to test its performance against the background-bias problem. The results are shown in Table. 3. We observe that, compared to the baseline (*Bsl*), the BUFF network 1) closes the performance gap between the *Origin* and *MB*, *RB* datasets. For example, when evaluating on *RB* dataset of CUHK-SYSU, the mAP of baseline drops by a larger margin than BUFF (20.4%(86.9-66.5) vs. 10.9%(92.2-81.3)). 2) achieves significant improvement on both the *MB* and *RB* datasets. For instance, on the *RB* of PRW, the mAP increases by 7.4% from 31.7% to 39.1%. The comparisons demonstrates the BUFF model can be more suitable for real-world applications where the testing scenarios have quite different background.

Table 4: Component analysis on PRW and CUHK-SYSU. We report the results evaluated with detected bounding boxes.

Method	PRW		CUHK-SYSU	
	mAP	rank-1	mAP	rank-1
Bsl + FAM	38.2	76.2	87.1	88.1
BUF_{c4}(bsl)	36.3	74.5	84.3	84.8
$BUF_{c3,4}$	41.0	78.8	89.5	90.3
$BUF_{c2,3,4}$	**42.6**	80.1	89.5	90.4
$BUF_{c1,2,3,4}$(BUF)	41.9	**80.2**	**89.8**	**90.7**

5.5 Component Analysis

Effect of FAM. We study the effect of FAM by conducting the *Bsl+FAM* experiment. The difference between *Bsl+FAM* and Bsl model is that *Bsl+FAM* adds a FAM on the top of the *conv4*. Thus the bounding box features for re-ID are pooled from the foreground-aware feature maps, i.e., the output of FAM. As shown in Table. 4, the *Bsl+FAM* (first row) achieves considerable performance gains over the baseline model, the results of which are shown in the second row. On CUHK-SYSU, the FAM brings improvement of 2.9% and 3.3% for mAP and rank-1, while on PRW, the mAP and rank-1 are improved by 1.8% and 1.7%, respectively.

Effect of BUF. In BUFF, we exploit features from several stages of the ConvNet for learning re-ID feature. Here, we conduct experiments to validate the effectiveness of features from different stages. In Table. 4, the $BUF_{c3,4}$ indicates the BUF exploits features of *conv3* and *conv4* for feature fusion. Thus BUF_{c4} represents the baseline (bsl) model. Compared to BUF_{c4}, the mAP of $BUF_{c3,4}$ increases by 3.7% from 36.3% to 41.0% on PRW and by 5.2% from 84.3% to 89.5% on CUHK-SYSU. When the features of *conv2* (the fourth row) and *conv1* (the last row) are combined, the rank-1 performance is further improved. We observe that the *conv*1 features bring a small margin of gains on the two datasets. That is because we follow the setting in OIM [34] to fix the parameters of *conv*1 layer. Moreover, as shown in Fig. 4, we compare the performances of single-level and the fused features. $BUF_{ci}(i \in 1, 2, 3)$ outperforms BUF_{c4} by a considerable margin, that indicates the potential to exploit the features of shallower layers for re-ID representation enhancement. The BUF (dashed line) achieves superior performance than BUF_{ci}, which validates the effectiveness of the bottom-up fusion mechanism.

5.6 Comparison with State-of-the-art Methods

In this section, we compare results of the proposed BUFF network with the state-of-the-art methods on PRW and CUHK-SYSU. The comparison methods are grouped into *task-joint* (denoted as *Y*) and *task-cascaded* (denoted as *N*). The *task-joint* group including OIM [34], IAN [32], NPSM [20], LCG [36] and QEEPS [23]. The *task-cascaded* methods adopt two separate steps of detection and re-ID. In this group, MGTS [3] uses VGGNet [27] based Faster R-CNN [25] detector, CLSA [16] employs vanilla Faster R-CNN, while RDLR [12] uses Faster R-CNN with FPN [19].

Evaluation on PRW. In Table 5, we report the results on PRW. We observe that 1) Compared to the *task-joint* methods, our methods exceeds the second best model QEEPS [23] by 5% mAP and 4.3% rank-1. The BUFF network achieves 81.0% rank-1, which sets new state-of-the-art results on PRW dataset. We believe that such performance gain benefits from mitigating the background-bias and

Table 5: PRW evaluation, where Y and N denotes *task-joint* and *task-cascaded* method, respectively. The shorter sides of input images are resized to 600 pixels, and the best performances are in bold.

Methods	Publication	Type	mAP	rank-1
OIM [34]	CVPR17	Y	21.3	49.9
IAN [32]	PR19	Y	23.0	61.9
NPSM [20]	ICCV17	Y	24.2	53.1
LCG [36]	CVPR19	Y	33.4	73.6
QEEPS [23]	CVPR19	Y	37.1	76.7
CLSA [16]	ECCV18	N	38.7	65.0
MGTS [3]	ECCV18	N	32.6	72.1
RDLR [12]	ICCV19	N	**42.9**	70.2
BUFF		Y	42.2	**81.0**

Table 6: Evaluation on CUHK-SYSU with gallery size of 100. Results of mAP (%) and top-1(%) accuracies are reported. The shorter sides of input images are resized to 600 pixels.

Methods	Publication	Type	mAP	rank-1
OIM [34]	CVPR17	Y	75.5	78.7
IAN [32]	PR19	Y	76.3	80.1
NPSM [20]	ICCV17	Y	77.9	81.2
RCAA [2]	ECCV18	Y	79.3	81.3
I-Net [15]	ACCV18	Y	79.5	81.5
QEEPS [23]	CVPR19	Y	84.4	84.4
LCG [36]	CVPR19	Y	84.1	86.5
MGTS [3]	ECCV18	N	83.0	83.7
CLSA [16]	ECCV18	N	87.2	88.5
RDLR [12]	ICCV19	N	**93.0**	**94.2**
BUFF		Y	90.6	91.6

enhancing the discriminative capability of re-ID representation. 2) The *task-cascaded* methods do not suffer from the background-bias problem, because the re-ID model takes the cropped pedestrian boxes as input. The BUFF network obtains comparable mAP with RDLR (42.2*vs*.42.9) and outperforms MGTS by 8.9%(81.0 − 72.1) in rank-1, which demonstrates the advantage of training detection and re-ID in a joint manner.

Evaluation on CUHK-SYSU. Table 6 shows the results on CUHK-SYSU with gallery size of 100. Our BUFF network outperforms most of the previous methods, including the *task-cascaded* methods, e.g., MGTS and CLSA. It is noted that the mAP and rank-1 accuracies of RDLR outperform all the compared methods, including the proposed BUFF, by a considerable margin. RDLR is a *task-cascaded* method that has a separated re-ID model, it is thus convenient to introduce existing techniques from re-ID domain, e.g., bag-of-tricks [22], to obtain a high-performance re-ID model. For examples, the mAP of their baseline is 8% better than ours on CUHK-SYSU (92.2% vs. 84.2%). However, the proposed BUFF network extracts re-ID features from the feature maps of the scene images rather than the bounding boxes images. Therefore, it is inflexible to apply those tricks, e.g., random erasing [41]. The performance gap between RDLR and BUFF is acceptable, because BUFF simplifies the training procedure and enjoys faster inference.

Comparison of efficiency. In Table. 7, we compare the number of model parameters and the average time taken by the models to process a single gallery image. The models are all evaluated with

Table 7: Comparison of efficiency on PRW. The input image size is set to 600×1000 (height $\times$ width).

Method	Parameters (M)	GPU	Time (sec)
Baseline	35.05	TITAN Xp	0.117
RDLR [12]	56.21	TITAN Xp	0.368
BUFF(ours)	37.01	TITAN Xp	0.127

ground truth bounding boxes. We observe that BUFF enjoys the advantages of saving model parameters and faster inference, which makes BUFF more practical than RDLR in realworld applications. These advantages benefits from two aspects. Firstly, the BUF subnet only introduce about additional 2M parameter while RDLR adopts two ResNet50 for detection and re-ID respectively. Secondly, BUFF simply use a conv combined with a FC layer as the RCNN while RDLR uses *conv5* of ResNet50. As a result, BUFF is nearly 3 times faster than RDLR (0.127 *sec* vs. 0.368 *sec*), and saves 19M of model parameters (56.21M−37.01M).

6 CONCLUSIONS

In this work, we conduct diagnostic analysis with the newly created background-influence datasets to study the influence of background on the *task-joint* baseline. Motivated by the study, we propose BUF to fuse large receptive field, semantically strong features with small receptive field, semantically weak features via a bottom-up pathway and transform layers. The BUF can be implemented flexibly in an end-to-end training framework. Moreover, we introduce FAM to guide the network to pay attention to the foreground regions for leanring discriminative visual representation. We show that it is a promising way to enhance the re-ID representation from a respect of feature pyramids, and the efficiency and simplicity of BUF will benefit future research of person search and related applications. In the future, we will explore how knowledge distillation can be use to enhance the capability of the BUFF network.

ACKNOWLEDGMENTS

This work is funded by the National Key Research and Development Program of China (Grant No.2016YFB1001005), the National Natural Science Foundation of China (Grant No. 61673375, No.61721004 and No.61876181), and the Projects of Chinese Academy of Science (Grant No. QYZDB-SSW-JSC006).

REFERENCES

[1] Garrick Brazil, Xi Yin, and Xiaoming Liu. 2017. Illuminating pedestrians via simultaneous detection & segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*. 4950–4959.

[2] Xiaojun Chang, Po-Yao Huang, Yi-Dong Shen, Xiaodan Liang, Yi Yang, and Alexander G Hauptmann. 2018. RCAA: Relational context-aware agents for person search. In *The European Conference on Computer Vision (ECCV)*. 84–100.

[3] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. 2018. Person search via a mask-guided two-stream cnn model. In *The European Conference on Computer Vision (ECCV)*. 734–750.

[4] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. 2019. Abd-net: Attentive but diverse person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*. 8351–8361.

[5] Navneet Dalal and Bill Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. IEEE Computer Society, 886–893.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee, 248–255.

[7] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. 2014. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 36, 8 (2014), 1532–1545.

[8] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. 2009. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 32, 9 (2009), 1627–1645.

[9] Ross Girshick. 2015. Fast r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*. 1440–1448.

[10] Yuan Gong, Yizhe Zhang, Christian Poellabauer, et al. 2019. Second-order Non-local Attention Networks for Person Re-identification. *arXiv preprint arXiv:1909.00295* (2019).

[11] Douglas Gray, Shane Brennan, and Hai Tao. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, Vol. 3. Citeseer, 1–7.

[12] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. 2019. Re-ID Driven Localization Refinement for Person Search. In *The IEEE International Conference on Computer Vision (ICCV)*. 9814–9823.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*. 2961–2969.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[15] Zhenwei He and Lei Zhang. 2018. End-to-end detection and re-identification integrated net for person search. In *Asian Conference on Computer Vision*. Springer, 349–364.

[16] Xu Lan, Xiatian Zhu, and Shaogang Gong. 2018. Person search by multi-scale matching. In *The European Conference on Computer Vision (ECCV)*. 536–552.

[17] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1, 4 (1989), 541–551.

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 152–159.

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2117–2125.

[20] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. 2017. Neural person search machines. In *The IEEE International Conference on Computer Vision (ICCV)*. 493–501.

[21] Chen Change Loy, Tao Xiang, and Shaogang Gong. 2009. Multi-camera activity correlation analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1988–1995.

[22] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[23] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. 2019. Query-guided End-to-End Person Search. *arXiv preprint arXiv:1905.01203* (2019).

[24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. 2017. In *Long Beach, California, USA: Autodiff Workshop*.

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeuralIPS)*. 91–99.

[26] Ergys Ristani and Carlo Tomasi. 2018. Features for multi-target multi-camera tracking and re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6036–6046.

[27] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[28] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *The European Conference on Computer Vision (ECCV)*. 480–496.

[29] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. 2018. Eliminating background-bias for robust person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5794–5803.

[30] Xiaogang Wang. 2013. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters* 34, 1 (2013), 3–19.

[31] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 79–88.

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

[33] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1249–1258.

[34] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2017. Joint detection and identification feature learning for person search. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3415–3424.

[35] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. 2014. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 937–940.

[36] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. 2019. Learning Context Graph for Person Search. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2158–2167.

[37] Shoou-I Yu, Yi Yang, and Alexander Hauptmann. 2013. Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3714–3720.

[38] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *The IEEE International Conference on Computer Vision (ICCV)*. 1116–1124.

[39] Liang Zheng, Yi Yang, and Alexander G Hauptmann. 2016. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984* (2016).

[40] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. 2017. Person re-identification in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1367–1376.

[41] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2017. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896* (2017).