End-to-End Thorough Body Perception for Person Search

Kun Tian,[1] Houjing Huang,[2] Yun Ye,[1] Shiyu Li,[1] Jinbin Lin,[1] Guan Huang[1]
[1]Horizon Robotics, [2]Institute of Automation, Chinese Academy of Sciences
{kun.tian, yun.ye, shiyu.li, jinbin.lin}@horizon.ai, houjing.huang@nlpr.ia.ac.cn, huangguan13@mails.ucas.ac.cn

Abstract

In this paper, we propose an improved end-to-end multi-branch person search network to jointly optimize person detection, re-identification, instance segmentation, and keypoint detection. First, we build a better and faster base model to extract non-highly correlated feature expression; Second, a foreground feature enhance module is used to alleviate undesirable background noise in person feature maps; Third, we design an algorithm to learn the part-aligned representation for person search. Extensive experiments with ablation analysis show the effectiveness of our proposed end-to-end multi-task model, and we demonstrate its superiority over the state-of-the-art methods on two benchmark datasets including CUHK-SYSU and PRW.

Introduction

Person search (Xiao et al. 2017; Zheng et al. 2017) aims to find a probe person in provided gallery images of the real world scenarios such as video surveillance for criminal search, and multi-camera multi-target tracking (Wen et al. 2017). Unlike person re-identification (ReID) where person is cropped from the original image and resized to a fixed scale, person search addresses the problem of identifying person without any prior knowledge of location and scale. Naturally, ReID accuracy will be affected by the detection result. However, what equally important but easily overlooked is that background clutter and misalignment can also lead to mismatching between query and gallery.

Background clutter refers to the perturbance of complex scene in the process of person feature expression, which is shown in Figure 1 (a). Another key challenge in person ReID is the misalignment caused by cropping errors and pose variations. For example, in Figure 1 (b), the box in column 1 represent legs, while the boxes in columns 2 and 3 represent the feet and background information, respectively. Furthermore, large pose variation result from camera views and person motion increase the difficult of matching, which is illustrated in Figure 1 (c).

Motivated by these observation, we explore how to more effectively integrate person detection and ReID into an end-to-end framework and make this multi-task model achieve better performance than the single-task ones. A significant gain from our baseline is observed over (Xiao et al. 2017). To the best of our knowledge, this is the first time that two sibling branches, namely person instance segmentation and keypoint detection branches, are introduced to person search model. We argue that by utilizing Foreground Feature Enhance Module (FFEM), the negative effect caused by background clutter can be mitigated. Furthermore, a keypoints-guided learning algorithm is introduced to deal with misalignment between images. All the above improvements enable our model a thorough body perception. Three major contributions are as follows:

- We build a better and faster base model, which is approximately two times faster than the original method while achieving a superior accuracy.
- We propose FFEM to effectively enrich the semantic feature of foreground person, which can smooth the response of CNN and reduce excessive attention to local areas.
- We put forward a feature learning algorithm to obtain part-aligned representation of person. Experimental results demonstrate that our Body Perception Network (BP-

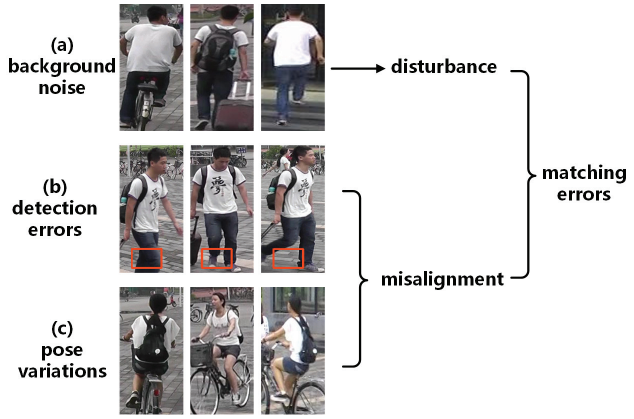


Figure 1: Three factors that may affect person search performance. (a) A person has similar posture but in different scenes. (b) There are some errors in the cropped output of detector, and the resulting negative impact can also be caused by occlusion. (c) A person is captured in similar scenes with three postures: back, front and side.

Net) surpasses all the current state of the arts on two benchmark datasets.

Related Work

Person Search

Xu et al. first proposed the concept of person search. Sliding window search method was used for detection and Fisher vectors were applied for matching. While the performance was limited due to manual extracted features and inefficient sliding algorithm. To improve search accuracy, Xiao et al. treated person detection and ReID as a joint optimization problem with a model based on Convolution Neural Networks. Zheng et al. proved that R-CNN based detection model with metric learning can yield significant ReID accuracy improvements. The deep learning method has become the mainstream of the research on person search. Apart from jointly optimization, Chen et al. and Xu et al. provided separately trained models which also achieved remarkable search accuracy. Recently, Yan et al. used a graph to learn similarities between target persons based on contextual information and Munjal et al. introduced a query-guided network to model global similarities between images.

Person Detection and Person Re-identification

Over the past few years, CNN-based person detectors such as (Zhang, Benenson, and Schiele 2017; Zhang, Yang, and Schiele 2018) have developed rapidly. In addition, Mask R-CNN proposed by He et al. unified detection and instance segmentation into a common model, which inspires us to explore the potential of jointly training person search model together with other body perception tasks.

Various CNN-based person ReID methods (Liu et al. 2017b; Xu et al. 2018; Xiao et al. 2016; Zheng et al. 2017) can be divided into two sets. The first set is based on feature representation. The goal of these models is to reduce the intra-class distance while increase the inter-class distance. The second set regard ReID as a classification problem. These methods have achieved good performance on the benchmarks, based on assumption that person images are accurately cropped. However, real-world applications produce imperfect detection, which will inevitably harm ReID performance due to background noise and misalignment. This problem causes regular spatial partition such as grid cell (Ahmed, Jones, and Marks 2015) and horizontal stripe (Sun et al. 2018) to be unreliable. As in the cases in Row 3 of Figure 1, one's legs will be matched to the background or feet region, which even deteriorate the accuracy.

Method

In this section, we firstly present an overview of our unified framework, and then explain how our proposed module and algorithm alleviate the influence from background clutter and feature misalignment.

Overview

In this paper, we propose a unified framework to obtain thorough body perception of person (BPNet). As illustrated in Figure 2, we use ResNet-50 (He et al. 2016) with FPN (Lin et al. 2017) as backbone, shared by all other sub-task networks, including person detection, instance segmentation, keypoint detection and re-identification.

In order to jointly train all four tasks in our framework, we use COCO pre-trained model to generate pseudo mask and keypoints annotations for CUHK-SYSU and PRW. Considering the domain shift between COCO and person search dataset, we believe that if there are manual annotations in the future, our proposed module and algorithm can further improve the accuracy of person search. We conduct extensive experiments to show that original person search model can be trained harmoniously with segmentation task as well as keypoint detection task. As far as we know, this is the first time that background clutter and feature misalignment is considered in person search task. We speculate the reason that there are few studies on the refined local representation is the size of person bounding box varies considerably, so the features within the same physical meaning area (i.e. arms, legs) differ greatly. In this regard, we adjust the structure of the original model and enable it to perceive person edge mask and body joints. Thanks to FFEM, which enhance global representation of target person but with negligible overheads, our model can extract features sensitive to the entire foreground area. Besides, part-based learning algorithm effectively deal with the problem that body parts are spatially misaligned. Briefly, our multi-task network can obtain fine-grained semantic information while reducing the misalignment problem through abstract part regions.

Loss Function

As elaborated above, the final model named BPNet uses a joint loss L to optimize its parameters during training stage, where L_{cls}, L_{box}, L_{mask}, L_{kp} and L_{reid} denote the loss function of all sub-tasks. In our experiments, loss weights λ_1 to λ_5 are designed to balance the training process. In this paper, the values of λ_1 to λ_4 are set to 1 and λ_5 to 0.2.

$$L = \lambda_1 L_{cls} + \lambda_2 L_{box} + \lambda_3 L_{mask} + \lambda_4 L_{kp} + \lambda_5 L_{reid} \quad (1)$$

For the loss function described above, L_{cls} and L_{mask} are calculated by Softmax loss and sigmoid-activated cross entropy, respectively. In addition, we treat bounding box refinement and keypoint detection as regression problems and losses are estimated by smooth L1 function. As for ReID branch, we replace the lookup table with the classification loss, in company with a circular queue (CQ) to predict the class-id for person identification. Corresponding formulas are as follows:

$$p_i = \frac{\exp(w_i^T x/\tau)}{\sum_{j=1}^{L}\exp(w_j^T x/\tau) + \sum_{k=1}^{Q}\exp(u_k^T x/\tau)} \quad (2)$$

$$q_i = \frac{\exp(u_i^T x/\tau)}{\sum_{j=1}^{L}\exp(w_j^T x/\tau) + \sum_{k=1}^{Q}\exp(u_k^T x/\tau)} \quad (3)$$

The Equation 2, 3 represent the probability that extracted person feature x belongs to the i-th labeled ID and i-th unlabeled ID respectively, where temperature τ can influence the distribution of different x in $\mathbb{R}^D$. Among the symbolic

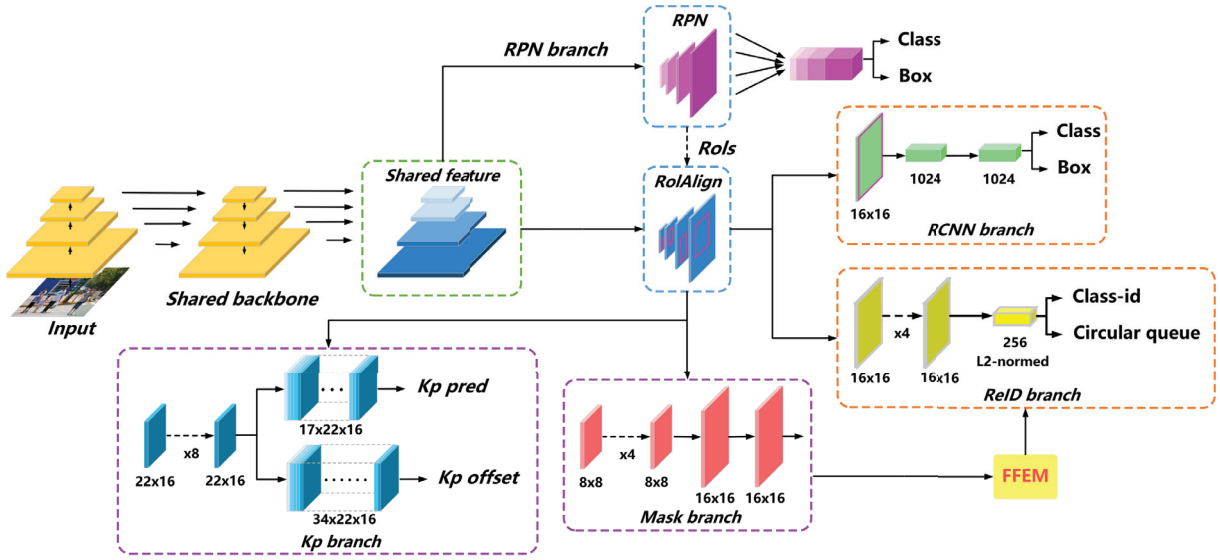
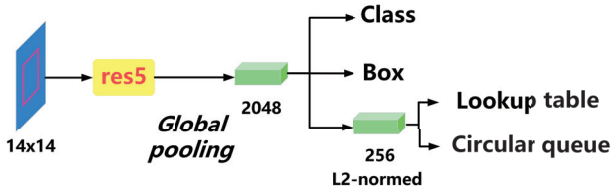


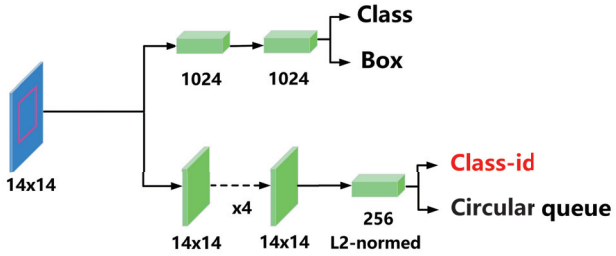
Figure 2: BPNet framework. We adopt FPN as our backbone to generate features shared by four parallel branches, namely *RCNN branch*, *ReID branch*, *Mask branch*, and *Kp branch*. This multi-task learning model can be mainly divided into three parts: backbone network for features extraction, light-head person search module which are framed by the orange dotted line, and auxiliary mask, keypoint branches that are framed by the purple dotted line. FFEM is the abbreviation of foreground feature enhancement module. Best viewed in color.

Method	CUHK-SYSU		PRW		Runtime(s)	
	Rank-1	mAP	Rank-1	mAP	C	P
heavy+OIM	78.7	76.5	42.7	22.1	0.30	0.37
light+Softmax	83.4	80.9	55.9	31.3	0.15	0.24

Table 1: Effects of model structure and loss function. "C" and "P" represent two datasets.



(a) heavy-head+OIM



(b) light-head+Softmax

Figure 3: The original model and our baseline

representation, $\mathbb{R}$ is the L2-normed feature space; D is the dimension of each identity feature; L is the number of nodes in the fully connected layer (FC); Q indicates the length of circular queue and w_i represents the parameter to be learned for the i-th node in FC. In addition, we denote the features stored in CQ by $u \in \mathbb{R}^D$. Finally, we use Softmax loss function to update FC parameters. Calculating the gradient of loss function with respect to x will use p_i and q_i.

Better and Faster Baseline

First, we re-implement (Xiao et al. 2017) using MXNet for fair comparison. The person search score is illustrated in Table 1. Our implementation (without newly added mask, keypoint branches) gets a slightly better results than original paper, probably because RoI Align is used instead of RoI Pooling to reduce quantization errors. The original model passes the pooled feature map to the rest conv4_4 to conv5_3 of the ResNet-50, followed by a global average pooling (GAP) layer to produce a feature vector. However, one notable point is that deeper network will increase the risk of overfitting and decrease the inference rate. We replace the last ResNet-block unit and subsequent GAP with four cascade convolution layers, so as to retain spatial information of person features and achieve dimensional reduction (1024-dim→256-dim). Compared with (Xiao et al. 2017), our light-head model achieves better accuracy and greater inference efficiency. Therefore, we regard it as our baseline and conduct subsequent research.

Foreground Feature Enhance Module

In complex scenarios, redundant scene information may be incorrectly weighted by person feature extractors and thus affect the following matching process. To deal with this

Figure 4: Two sets of comparisons. The 1st and 4th graphs depict the detected person. The 2nd, 5th and 3rd, 6th images show the heatmap of person features extracted from ReID branch without/with FFEM, respectively. The warmer the color, the stronger the response to the region. Therefore, it can be concluded that FFEM produces a comprehensive attention map, thus focusing on the holistic person rather than overfitting local information such as clothing color/texture.

problem, human body mask are used to alleviate the influence from background clutter. Considering the complementary relationship between ReID branch and mask branch, we design FFEM to preserve and enhance the spatial information in person feature map, so as to improve the expressive ability of feature representation in ReID branch. In other words, global feature alignment and weighted fusion are performed in FFEM, which is illustrated in Figure 4. Additionally, Figure 5 shows the training and inference stages of FFEM.

$$R \circ M = (R_{jk} \cdot M_{jk}) = \begin{pmatrix} r_{11} \cdot m_{11} & \cdots & r_{1k} \cdot m_{1k} \\ \vdots & \ddots & \vdots \\ r_{j1} \cdot m_{j1} & \cdots & r_{jk} \cdot m_{jk} \end{pmatrix} \tag{4}$$

$$ER_i = R_i \circ M_i \oplus R_i \tag{5}$$

For training stage, we broadcast the corresponding pseudo labels to have the same shape with the feature of each person RoI proposal in ReID branch. The mask feature fusion process formulated in Equation 4, 5. First, we extract foreground sensitive feature by applying Hadamard Product between global feature maps and human masks. Here $R_i \in \mathbb{R}^{C \times H \times W}$ denotes an input feature map of i-th person proposal, M_i represents the broadcasted mask label and $\circ$ denotes the Hadamard Product operator which performs element-wise product on two matrices or tensors. Second, considering that some foreground information may be lost when using human mask to filter background noise, we perform element-wise addition as shown in the Equation 5. Finally, enhanced feature maps ER are followed by 4 cascade convolution layers and a fully connected layer to produce identity-related global feature vector f_g. At the inference stage, we directly use the predicted probability map and binarize it with a threshold of 0.5. The subsequent process is consistent with the training phase. Based on the above, FFEM prevents model from mistaking foreground areas for background noise and preserves high-quality global context of extracted features, which also highlight the useful region in person bounding box.

Keypoints-Guided Learning Algorithm

Although FFEM alleviates the perturbance of background noise, global context features suffer from body part misalignment between two persons. Motivated by this observation, keypoints estimation is introduced as an auxiliary task to align part representation in a multi-task learning manner.

Algorithm 1: Part-aligned feature learning algorithm

Input: y-coordinates of person $\{y_{mid}, y_{qua}, y_{min}, y_{max}\}$, keypoints information $\{y_*(e.g.\ y_{hips}, y_{shoulders}), k_{vis}\}$ and shared backbone features$\{F\}$

Output: part-aligned feature$\{F_p\}$,part-visibility$\{V_p\}$

for $p = 0; p \leq P;$ **do**

  if $p = 0$ **then**

    $p_{min} = y_{min}\ \ p_{max} = max(y_{mid}, y_{hips})$;

  else if $p = 1$ **then**

    $p_{min} = y_{shoulders}\ \ p_{max} = y_{hips}$;

  else if $p = 2$ **then**

    $p_{min} = y_{min}$;

    $p_{max} = max(y_{qua}, \frac{1}{2}(y_{hips} + y_{shoulders}))$;

  $f_p = roi\_align((p_{min}, p_{max}), \{F\})$;

  if $num(p_{min} < y_*[k_{vis}] < p_{max}) > TH$ **then**

    $v_p = 1$;

  else

    $v_p = 0$;

  $f_p \in \{F_p\}, v_p \in \{V_p\}$;

return $\{F_p\}$,$\{V_p\}$;

Following the definition in COCO dataset (Lin et al. 2014), 17 keypoints of human body are utilized to supervise the training of keypoint branch. Then, we propose a keypoints-guided learning algorithm to generate abstract region with the same physical meaning in different viewpoints, namely upper body, torso and head-shoulder. As shown in Figure 6, each region is defined by a neat rectangle, which does not refer to the precisely segmented body part, but to an abstract region with representational capability. More concretely, the process of the algorithm can be divided into three folds. First, get keypoints information. For each point to be detected, it is considered invisible if its position is outside the detected region or if it is obscured. In inference, we use confidence score to represent the visibility instead. Second, bounding box is divided into *P* parts, and the value of *P* is not greater than 3. The specific process of outputting the feature expression and visibility of each region is shown in Algorithm 1, where y_* and k_{vis} denotes the ordinates and visibility of 17 points. For balance training, TH as a threshold is usually set to 4. Last, each part learns discriminative embeddings respectively, under the supervision of ID labels. During training, model automatically determines whether the corresponding part is occluded or not according to the number of keypoints visible in this part. If a part defined above is occluded, then its loss will not be backpropagated. When testing, *P* part embeddings $\{F_p\}$ are further concatenated with the above f_g to produce aligned feature vector f

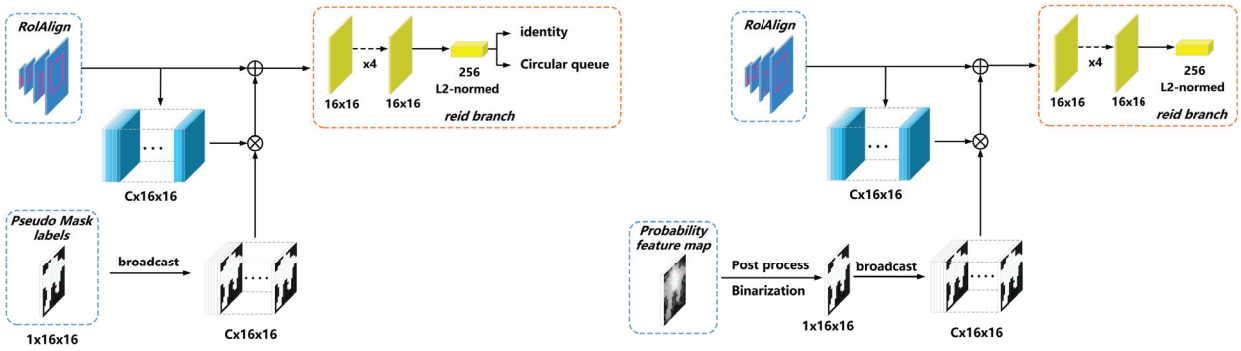


Figure 5: Module architecture. We illustrate how the proposed foreground feature enhance module (FFEM) works during training and inference stages. Specifically, the left sub-graph represents the training phase: pseudo mask label is used to enhance the feature map of ReID branch; the right sub-graph represents the inference stage: predicted probability map is fused with the feature map of ReID branch.

Figure 6: Blue joints corresponding to head-shoulder part, yellow joints corresponding to torso, the union of two colors joints corresponding to the green boundary of upper body, and the red border includes the whole body.

for re-identification. In summary, our learning algorithm has two main advantages. First, this algorithm aims to produce well-aligned parts descriptor so that the learned feature can be matched correctly for the same person under intensive pose changes. Second, since the visibility of abstract part is considered in training, our model is also robust to occlusion.

Experiment

Datasets and Evaluation Protocol

CUHK-SYSU contains 11,206 images with 5,532 labeled person IDs for training, 2,900 queries and a total of 6,978 gallery images for testing. Persons in this dataset vary largely in viewpoints, background and lighting conditions. PRW consists of 11,816 annotated video frames, 933 labeled identities, and 43,110 person bounding boxes. The training set contains 5,704 images with 483 labeled person IDs. For the test set, there are 2,057 probe person along with a gallery set of 6,112 whole scene images. We adopt the mean Average Precision (mAP) and the Cumulative Matching Characteristic (CMC Rank-k) as evaluation metrics. The gallery size of two datasets is set to 100 and 6,112 respectively.

Implementation Details

As well elaborated in Method Section, our approach is a unified model and can be trained in an end-to-end manner. We employ the ImageNet-pretrained (Russakovsky et al. 2015) ResNet-50 as backbone. The model is trained under the MXNet framework and each mini-batch has 4 or 3 images on one GTX 1080Ti due to the memory limitation. In details, the network is optimized for 23 epochs using mini-batch stochastic gradient descent with a weight decay of 0.00004 and a momentum of 0.9. We adopt the warm-up strategy (Goyal et al. 2017) to change the learning rate for the first 5 epochs and divide it by 10 at 19-th and 22-th epoch. In addition, we fix the batch normalization (Ioffe and Szegedy 2015) layers in the backbone while keep others trainable in sub-branches. The temperature scalar τ in Equation 2, 3 is set to 0.05. For CUHK-SYSU and PRW, the size of the Circular Queue is set to 5,000 and 500 respectively. The scale of the training and testing images is fixed to 600×1000 pixels unless otherwise noted and the training sample are also horizontally flipped in random.

Ablation Study

According to the analysis in Method Section, we add mask, keypoint branches to the base model. Our experiments demonstrate the combination of FFEM, learning algorithm and multi-task branches can increase the overall performance in both datasets, which is displayed in Table 2.

Specifically, the proposed FFEM effectively strength the foreground information in ReID feature map, which is conducive to extracting identity-sensitive features and further improving the accuracy of the person search model. Considering the newly established mask branch and feature fusion module, Rank-1/mAP are increased by 1.8%/2.2% and 4.4%/2.9%, respectively in CUHK-SYSU and PRW. Besides, we find that FFEM naturally achieves global feature alignment is another bonus, which also explains why FFEM can achieve a steady improvement on two datasets.

In addition, our proposed learning algorithm can effectively deal with the problem of feature misalignment. Since upper body region includes more identity information than lower body, such as clothing color, backpack and other personal items, we speculate that it is focusing on appropriate local features that enhance the descriptor. But for head-shoulder, the integration of this region introduces more noise

Method	FPN	Mask	FFEM	Kp	Algorithm	Scale aug	CUHK-SYSU		PRW	
							Rank-1	mAP	Rank-1	mAP
light+Softmax							84.1	81.0	56.3	31.2
light+Softmax	✓						85.0	82.1	59.6	34.2
light+Softmax	✓	✓					85.3	83.1	60.7	35.3
light+Softmax	✓	✓	✓				86.8	84.3	64.0	37.1
light+Softmax	✓	✓	✓	✓			86.5	84.0	62.0	36.9
light+Softmax	✓	✓	✓	✓	✓		88.7	86.4	66.7	40.4
light+Softmax	✓	✓	✓	✓	✓	✓	90.5	88.4	68.9	42.9

Table 2: Ablation study on our base model. "FPN" implies adding Feature Pyramid Network to the backbone. "Mask" and "Kp" represent that newly added branches. "FFEM" is our proposed module. "Algorithm" indicates that part-aligned features are taken into account. "Scale aug" means smaller multi-scale training. The short edge of the training image is increased from 416 pixels to 608 pixels in 32 pixels steps. Models are evaluated on CUHK-SYSU and PRW.

due to inaccurate partition, which deteriorates the overall system performance. Nevertheless, the ReID performance is still improved when compared with base model, which is because the combination of global representation and other two aligned local features reduces the impact of inaccurate estimate and information loss caused by head-shoulder region. As shown in Table 2, (1.9%/2.1%) and (2.7%/3.3%) gain for rank-1/mAP on two datasets compared to the method without using part-aligned feature. In conclusion, our algorithm is also able to differentiate body regions and the differentiation is adaptive to each input image for translation/pose invariance.

Figures 7 displays the effect of hyper-parameter P on model performance. For fair comparison, the dimensions of the final feature vector f without using our algorithm is set as $256 * (P + 1)$. We have more observations about the generated abstract parts. 1) Only increasing the dimension of the final feature vector makes model easy to overfit and the search accuracy decreases. But the model using our learning algorithm can alleviate this problem, i.e. adding torso region(dimension changes from 512 to 768) improves the accuracy of model. 2) The head-shoulder region is not suitable for analysis as an independent part, because the face is not frontal and with low resolution, thus accordingly not reliable for distinguishing different persons. In other words, after introducing head-shoulder part, the increased background noise is actually more than effective foreground information. 3) Our algorithm has the same variation trend on the two datasets, which proves its robustness. According to experimental results, we set P equal to 2 for the best performance. It is possible that in other datasets the optimal P obtained through validation is different.

Two groups of qualitative results are shown in Figure 8. The first group of pictures on the left side of dashed line indicates that for a same query, our base model and the model using FFEM, learning algorithm have dissimilar search results for the same gallery image. The base model produced false match predictions due to clothing similarity and viewpoint variability between persons. From the second group of pictures on the right side of the dashed line, we can see that the base model fails to find the true positive because target person is occluded by bicycles in the scene. In conclusion,

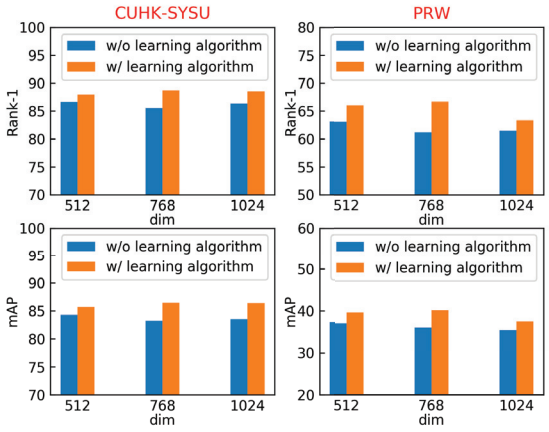


Figure 7: Performance of final representation on CUHK-SYSU and PRW with different dimension and corresponding P. "512" indicates addition of upper body, "768" and "1024" represent additional torso and head-shoulder regions, respectively. The values of all blue bars are lower than those corresponding to the 5th row of Table 2, the feature dimension of which is only 256-dim.

using FFEM and the learning algorithm can not only improve the response of our model to the foreground, but also have better robustness to occlusion and pose variations.

Comparison with State of the Arts

In this section, we compare our BPNet with previous state-of-the-art methods on CUHK-SYSU, PRW.

When the scale of test image is 600×1000 pixels, our BPNet achieves 90.5% Rank-1 and 88.4% mAP, surpassing the best end-to-end method QEEPS by 6.1% Rank-1 and 4.0% mAP respectively, which even outperforms two-stage model CLSA by 2.6% in Rank-1 and 1.9% in mAP. Moreover, we evaluate our model under different gallery size. Figure 9 depicts how mAP changes with different gallery size of [50,100,500,1000,2000,4000]. It can be seen that with the increase of gallery size, the performance of all models will

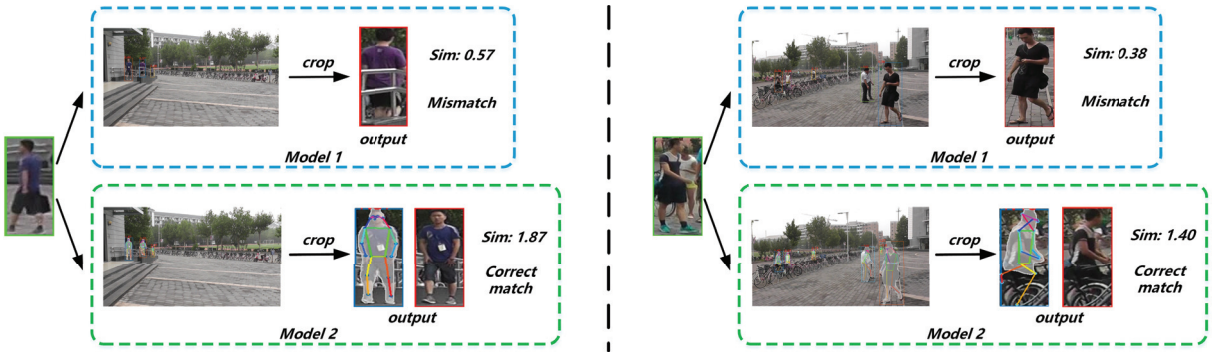


Figure 8: Matching results of two query persons. *Model 1* is our base model and *Model 2* use FFEM and the learning algorithm, which enables it to have a more comprehensive perception of the human body. Additionally, *Model 2* calculates the local similarity of upper body and torso region, and then adds the global similarity of the whole body as the final judgment basis.

Method	Rank-1(%)	mAP(%)
OIM(Xiao et al. 2017)	78.7	75.5
IAN(Xiao et al. 2019)	80.1	76.3
NPSM(Liu et al. 2017a)	81.2	77.9
CNN_v+MGTS*(Chen et al. 2018)	83.7	83.0
CGRL(Yan et al. 2019)	86.5	84.1
CLSA(Lan, Zhu, and Gong 2018)	88.5	87.2
QEEPS*(Munjal et al. 2019)	89.1	88.9
Our BPNet	90.5	88.4

Table 3: Evaluation on CUHK-SYSU based on CNN models. (*) indicates models use larger images with a sizes of 900×1500 pixels. Following tables share this annotation.

Method	Rank-1(%)	mAP(%)
OIM(Xiao et al. 2017)	49.9	21.3
IAN(Xiao et al. 2019)	61.9	23.0
NPSM(Liu et al. 2017a)	53.1	24.2
CLSA(Lan, Zhu, and Gong 2018)	65.0	38.7
Our BPNet	68.9	42.9
CNN_v+MGTS*(Chen et al. 2018)	72.1	32.6
CGRL(Yan et al. 2019)	73.6	33.4
QEEPS*(Munjal et al. 2019)	76.7	37.1
Our BPNet	87.9	48.5

Table 4: Evaluation on PRW based on CNN models. The 1st and 2nd highest scores are shown on the last and penultimate lines, respectively.

be degraded. However, BPNet keeps the highest score over other models under different gallery sizes and its degradation is smaller. This shows that our method is more robust against the gallery size.

For PRW, all known latest methods are not comparable due to the lack of unified evaluation criteria. One is provided by (Chen et al. 2018). For a query image in test set, all labeled person in the rest 6,111 images form its gallery set. The other criterion, considering the evaluation of multi-camera ReID task, filters out person who has same ID and is in the same camera as the query, and person whose ID is labeled as -2 from the gallery set, which improves the difficulty of search task. As indicated in Table 4, we test BPNet with both criteria. The evaluation process of methods above the horizontal line is unknown and we use the latter criterion for fairness, while methods below the horizontal line using the same test code (the former criterion). Anyway, our model outperforms all previous state of the arts, including end-to-end and two-stage approaches.

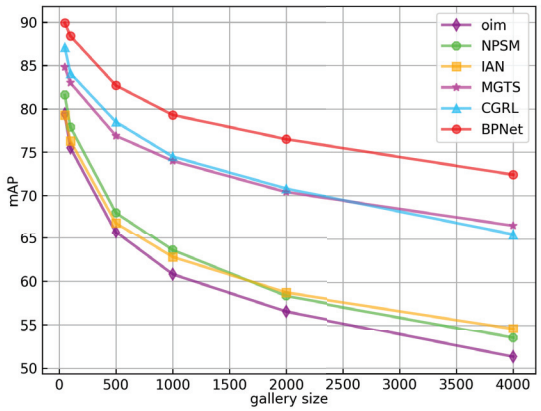


Figure 9: mAP comparison on CUHK-SYSU with varying gallery sizes. There is a considerable gap between the other approaches and our BPNet.

Comparison of Running Time

Limited by GPU memory, we do not use larger input images for training. But in order to fairly compare runtime with other state of the arts, we expand the size of test image to 900×1500 pixels. Considering these methods use different GPUs during testing, we also report TFLOPs. The comparison results are shown in Table 5. Besides, it is worth noting that reducing the number of proposals to 100 does not compromise mAP(88.3%) or Rank-1(90.3%) in search task, while the test speed is doubled. If not specified, the number of person proposals in other tables is set to 300.

Method / proposals	Runtime(s)	GPU(TFLOPs)
CNN_v+MGTS / 300	1.3	K80(8.7)
QEEPS / 300	0.3	P6000(12.0)
BPNet / 300	0.5	GTX 1080Ti(3.8)
BPNet / 100	0.3	GTX 1080Ti(1.8)

Table 5: Comparison of model running time when the size of test image is 900×1500 pixels.

Conclusion

In this paper, we present a novel thorough body perception framework and explicitly address background noise and the misalignment problem in person search task. The key factors that contribute to the superior performance of our approach are as follows. (1) Better and faster baseline. (2) FFEM enable global appearance feature map to provide a robust foreground representation under clutter environment. (3) The learning algorithm aims to partition the human body into abstract region instead of grids or strips, and thus is more robust to pose changes and different spatial distributions in human bounding box. We validate the effectiveness of our approach by demonstrating its superiority over the state-of-the-art methods on the standard benchmark datasets, and our contributions are complementary to each other.

References

Ahmed, E.; Jones, M. J.; and Marks, T. K. 2015. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*.

Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; and Tai, Y. 2018. Person search via a mask-guided two-stream cnn model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 734–750.

Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; and He, K. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Lan, X.; Zhu, X.; and Gong, S. 2018. Person search by multi-scale matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 536–552.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Liu, H.; Feng, J.; Jie, Z.; Jayashree, K.; Zhao, B.; Qi, M.; Jiang, J.; and Yan, S. 2017a. Neural person search machines. In *Proceedings of the IEEE International Conference on Computer Vision*, 493–501.

Liu, H.; Feng, J.; Qi, M.; Jiang, J.; and Yan, S. 2017b. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing* 26(7):3492–3506.

Munjal, B.; Amin, S.; Tombari, F.; and Galasso, F. 2019. Query-guided end-to-end person search. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3):211–252.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, 480–496.

Wen, L.; Lei, Z.; Chang, M.-C.; Qi, H.; and Lyu, S. 2017. Multi-camera multi-target tracking with space-time-view hyper-graph. *International Journal of Computer Vision* 122(2):313–333.

Xiao, T.; Li, H.; Ouyang, W.; and Wang, X. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1249–1258.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3415–3424.

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

Xu, Y.; Ma, B.; Huang, R.; and Lin, L. 2014. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Proceedings of the 22nd ACM international conference on Multimedia*, 937–940. ACM.

Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; and Ouyang, W. 2018. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2119–2128.

Yan, Y.; Zhang, Q.; Ni, B.; Zhang, W.; Xu, M.; and Yang, X. 2019. Learning context graph for person search. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, S.; Benenson, R.; and Schiele, B. 2017. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3213–3221.

Zhang, S.; Yang, J.; and Schiele, B. 2018. Occluded pedestrian detection through guided attention in cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6995–7003.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1367–1376.