# Bi-directional Interaction Network for Person Search

Wenkai Dong[1, 3], Zhaoxiang Zhang[1, 2, 3]*, Chunfeng Song[1, 3], Tieniu Tan [1, 2, 3]*
[1] Center for Research on Intelligent Perception and Computing, NLPR, CASIA
[2] Center for Excellence in Brain Science and Intelligence Technology, CAS
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences
{dongwenkai2016, zhaoxiang.zhang}@ia.ac.cn, {chunfeng.song, tnt}@nlpr.ia.ac.cn

## Abstract

*Existing works have designed end-to-end frameworks based on Faster-RCNN for person search. Due to the large receptive fields in deep networks, the feature maps of each proposal, cropped from the stem feature maps, involve redundant context information outside the bounding boxes. However, person search is a fine-grained task which needs accurate appearance information. Such context information can make the model fail to focus on persons, so the learned representations lack the capacity to discriminate various identities. To address this issue, we propose a Siamese network which owns an additional instance-aware branch, named Bi-directional Interaction Network (BINet). During the training phase, in addition to scene images, BINet also takes as inputs person patches which help the model discriminate identities based on human appearance. Moreover, two interaction losses are designed to achieve bi-directional interaction between branches at two levels. The interaction can help the model learn more discriminative features for persons in the scene. At the inference stage, only the major branch is applied, so BINet introduces no additional computation. Extensive experiments on two widely used person search benchmarks, CUHK-SYSU and PRW, have shown that our BINet achieves state-of-the-art results among end-to-end methods without loss of efficiency.*

## 1. Introduction

Person search [28] aims at localizing a target person in a gallery of unconstrained scene images. Compared with person re-identification (Re-id), it contains the process of generating person proposals from scene images, which makes it more suitable for real-world applications, such as video surveillance and security, video retrieval, and human-computer interaction. It is a challenging problem because of raw unrefined detections, camera view changes, low resolution, background clutter, and occlusion, *etc*.

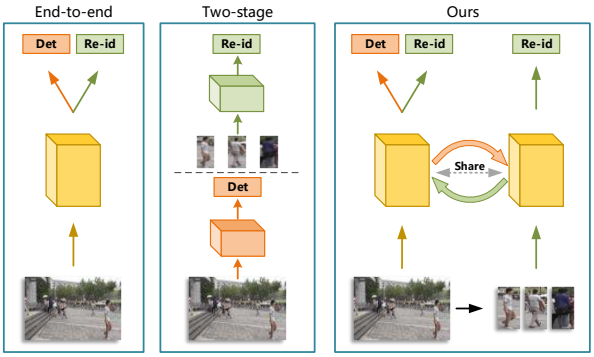



Figure 1. Comparison of three methods for person search. (a). Existing end-to-end framework. (b). Existing two-stage framework. (c). Our framework.

Person search is a fine-grained task which needs accurate human appearance information. Compared with the classification task in generic object detection, person search is more likely to be affected by the redundant context information outside the bounding boxes. For example, in the training procedure, the model discriminates some identities based on the various context rather than human appearance. During inference, the model may still fail to focus on persons, so the identity features lack the capacity of discrimination.

Existing approaches to person search separate this task into generating person proposals from scene images and person Re-id. As shown in Figure 1, they tackle two sub-tasks together in a multi-task framework (end-to-end methods) or separately via two independent networks (two-stage methods). For the end-to-end methods, existing works [26, 21, 24] apply multi-task frameworks based on Faster R-CNN to solve the two sub-tasks together. Similar to the pipeline of Faster R-CNN, an RoI-Pooling layer is applied to pool feature maps of the same size for each proposal. Generally, receptive fields in deep CNN such as ResNet

*Corresponding Author.

[12] are large, which means that the pooled feature maps of the proposals involve the context information outside the bounding boxes. Although some useful context is important for precise detection, yet there is still a large amount of irrelevant context in the surroundings. As aforementioned, such irrelevant context information can harm the person search performance. Moreover, despite achieving state-of-the-art performance, QEEPS [21] is a query-guided method, so it has to re-calculate proposals for different queries, which makes it impractical in the real world.

In the two-stage methods [15, 3], the training data of the Re-id models is a set of person patches cropped from scene images. Thus, the context information has little influence. However, two-stage methods may lead to a sub-optimal problem because the detection network and person Re-id model are trained separately.

The above observations motivate us to improve the existing end-to-end person search methods in the following aspects: (1) The burden of the redundant context information should be alleviated so that the model can learn more accurate human appearance information; (2) The model needs the guidance of person patches to discriminate human appearance information from redundant context information; (3) The method should be efficient at the inference stage, i.e., it is not query-guided.

Following the above argument, we propose a simple yet effective model, named Bi-directional Interaction Network (BINet), to learn more discriminative representations for persons. Inspired by the previous work [4], BINet consists of two branches, taking as inputs scene images and person patches cropped from them, respectively. In this way, the model has access to the person patches without context information outside the bounding boxes, which can help the model discriminate identities according to their appearance information. Moreover, we propose two interaction losses to achieve bi-directional feature-level and prediction-level interaction between two branches. The losses force the model to respond consistently to the paired data at two levels. We share the parameters between two branches. Thus, the additional branch can be removed at the inference stage and no extra computation is introduced.

We conduct extensive experiments to demonstrate the effectiveness of our proposed BINet, which boosts the person search performance by making the model focus more on persons than context. The contributions of this paper are three-fold:

- We propose the Bi-directional Interaction Network which can learn to focus on persons in the scene with the guidance of the cropped person patches.
- We design two interaction losses to perform bi-directional interaction between the branches during the backward process. The interaction can make the model learn more discriminative identity representations.
- Our BINet brings significant performance improvements on popular benchmarks without additional parameters or computation. In particular, compared with our baseline [26], it achieves improvements of 3.6% on CUHK-SYSU and 10.3% on PRW in mAP accuracy.

## 2. Related Work

**Person search.** Person search [28] aims to localize a target person in a gallery of whole scene images. Many methods have been proposed to solve this problem since the publication of two large scale datasets, CUHK-SYSU [26] and PRW [36]. These methods separate the task into two parts, person proposals generation and identity matching, and solve them separately or jointly. For two-stage methods, they consider person search as a combination of pedestrian detection [9, 7, 30, 22, 8, 32, 33, 23] and person Re-id [35, 18, 34, 14, 31, 17, 5, 25] and solve them with two separate models. For example, Lan *et al.* [15] and Chen *et al.* [3] apply Faster R-CNN to detect person proposals from scene images and then train a person Re-id model to solve this problem. More specifically, Lan *et al.* [15] identify the multi-scale matching problem caused by the detector and exploit knowledge distillation to address this problem. Chen *et al.* [3] extract more representative features for each person by a two-stream model. Han *et al.* [11] introduce a ROI transform layer to jointly optimize the two networks.

Other works choose to solve the two sub-tasks in an end-to-end fashion. For example, in [26, 24, 29, 21], they all develop an end-to-end person search framework based on Faster R-CNN [23] to jointly handle the two aspects. Instead of detection-based methods, other methods [19, 2] recursively shrink the search region to more accurately locate the target person in the scene with the guidance of the information of the query. Liu *et al.* [19] propose Conv-LSTM [27] based Neural Person Search Machines (NPSM) to perform the search process. Chang *et al.* [2] make the search process as a conditional decision-making process and introduce deep reinforcement learning to the field of person search.

**Influence of context on instance representation learning.** In Faster-RCNN, the inputs to the end-to-end network are scene images and feature maps for each proposal are pooled with an RoI-Pooling/Align layer. Due to the large receptive fields in deep CNN, these feature maps involve a large amount of context information outside the region of interests. It is commonly believed that context information is important for precise detection [4]. However, irrelevant context introduced by the large receptive fields may lead to wrong classification results. Therefore, Cheng *et al.* [4] train a separate RCNN [10] for classification, named DCR

module, which takes as inputs a mini-batch of RoIs sampled from full images, and merge the classification scores of DCR module and Faster R-CNN to obtain the final classification scores. Inspired by this and feature mimicking [1, 13, 16], Zhu *et al.* [37] also apply a similar framework, where the two branches share parameters, and introduce a mimicking loss to force the model focus on the objects. For person search, although query-guided methods [2, 21] can select helpful context information with the guidance of queries, they need to re-calculate proposals for different queries, which makes them impractical in real-world applications.

# 3. Bi-directional Interaction Network

## 3.1. End-to-end Framework for Person Search

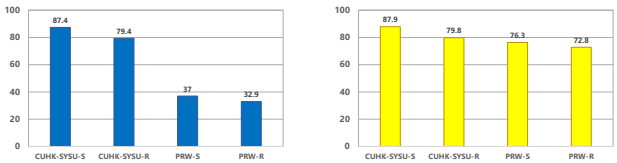
As aforementioned, we aim to solve the person search in an end-to-end network. Therefore, we take this multi-task network [26] based on Faster R-CNN as our baseline and improve it to make it more efficient. The overall framework of this baseline is illustrated in Figure 1.

Specifically, as shown in Figure 3, we adopt ResNet-50 [12] as our backbone network and separate it into two parts. The first part (conv1 to conv4) processes scene images and outputs 1024-channel feature maps. Then a region proposal network (RPN) is built on these feature maps to generate region-of-interests (RoIs). After non-maximum suppression, we keep 128 RoIs and exploit RoI-Align to pool a $1024 \times 14 \times 6$ region from the stem feature maps for each RoI. Then these RoIs are passed through the second part (conv5) of ResNet-50, followed by a global average pooling layer. Finally, the features are fed into three branches. Following the previous works, we adopt the Online Instance Matching (OIM) loss [26] to supervise feature learning for each identity. For the details of OIM loss, please refer to this work [26].

Compared with the original end-to-end network for person search, we reduce the layers in the second part (conv4_4 to conv5_3 in [26]) and modify the output size ($14 \times 14$ in [26]) of the RoI-Align layer. Since the aspect ratios of the annotated bounding boxes mostly range from $0.5$ to $0.25$, it is reasonable to modify the output size of the RoI-Align layer to $14 \times 6$. Through the above modifications, we reduce the overall computation cost, so the model is more efficient during inference.

## 3.2. Problems with the End-to-end Framework

Although the baseline method can handle the person search task in an end-to-end fashion, yet there is a main drawback in this method, which can be a bottleneck of the performance. To generate feature maps for each proposal, an RoI-Pooling/Align layer is applied to pool a region from the stem feature maps. Due to the large receptive fields in deep CNN, these feature maps contain redundant context information outside the bounding boxes. Although context information is helpful for the precise localization, however, for identity matching, which needs fine-grained appearance information, it can make the learned features less discriminative.




(a) mAP (b) top-1

Figure 2. Influence of the context information on the search performance on CUHK-SYSU and PRW datasets. The notations "S" and "R" denote the identity features are extracted from the scene images and cropped person patches, respectively. The former contains context information outside the bounding boxes while the latter only contains information inside the boxes.

To investigate the influence of context, we train the baseline model with scene images as [26] and report the results under two different evaluation settings in Figure 2. To exclude the influence of detection, the model is tested with ground truth RoIs on both datasets. The first test setting (CUHK-SYSU-**S** and PRW-**S**) is the same as [26], i.e., the trained model takes as inputs scene images and the features for proposals are pooled from the stem feature maps. In the second setting (CUHK-SYSU-**R** and PRW-**R**), we remove the detection-relevant parts from the baseline model (illustrated as the upper branch in Figure 3). The inputs to the network are the person patches cropped from scene images using ground truth RoIs. Thus, the features of proposals contain no context outside the bounding boxes. The results of mAP and top-1 are shown in blue and yellow, respectively. We observe that using cropped patches harms the performance significantly on both datasets. We conjure that in the training process, the model discriminates some identities based on different context information outside the bounding boxes rather than accurate appearance information. During inference, the identity features extracted from scene images are not discriminative enough because the model still discriminates persons based on the context to some degree. Therefore, we think that to learn more discriminative features, the model needs to pay attention to human appearance rather than the context.

## 3.3. Instance-aware Branch

In [21], the authors design a query-guided Siamese network to leverage information from query and gallery images and achieve state-of-the-art performance. With the guidance of queries, the model can focus on query-relevant information from gallery images. However, such query-

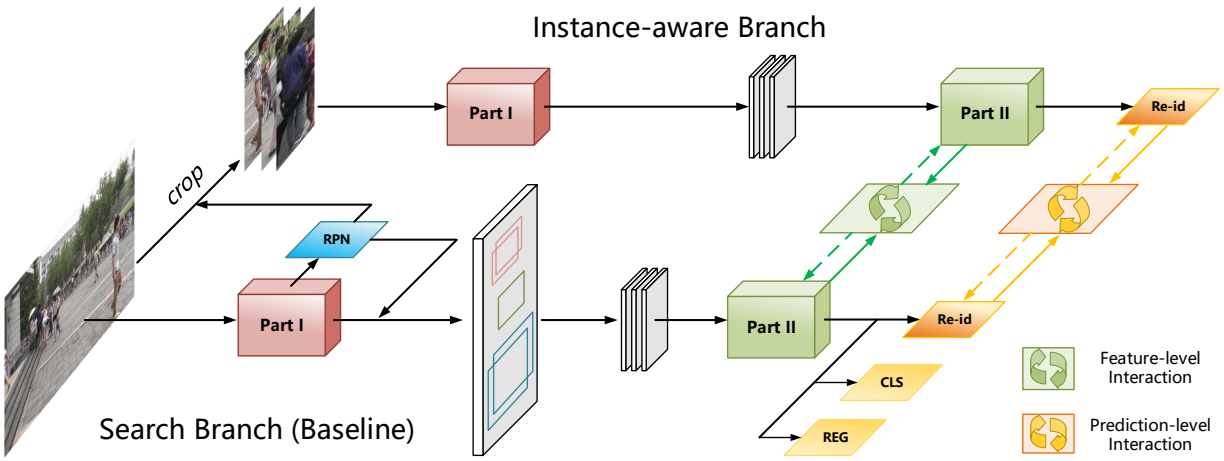



Figure 3. Our proposed framework. BINet takes as inputs scene images and cropped person patches. The common parts of the two branches **share parameters**. Bi-directional interaction between two branches is achieved by the interaction losses. During inference, we only apply the search branch. The dashed lines represent the directions of the gradients.

guided methods are inefficient in inference, i.e., the proposals generated from scene images are query-relevant, so for a new query, the proposals need to be re-computed. We argue that the model needs new guidance in the training procedure instead of queries.

Therefore, we turn to the patches cropped from input scene images to help the model learn to discriminate identities based on human appearance. Specifically, as shown in Figure 3, besides the major branch, an additional branch, named instance-aware branch, is added for person Re-id training. For each positive RoI $b$ fed into the second part in the search branch, the person patch corresponding to it is cropped and resized to $224 \times 96$ pixels. In the instance-aware branch, the backbone network operates on the resized cropped patches and outputs 2048-dimension features after the second part. Note that the output feature maps of the first part are of size $14 \times 6$, which is the same as the output size of the RoI-Align layer, so there is no RoI-Align layer in this branch. After that, an embedding layer of 256-D is applied, producing feature representations for each person. Finally, the OIM loss is applied to supervise feature learning. Since the inputs to this branch are cropped patches, the features contain no context outside the bounding boxes. The common parts in both branches share parameters. In this way, the model has access to the "clean" data directly and can learn to focus on the person. Moreover, during inference, only the major branch is applied, so no additional parameters and computation are introduced.

### 3.4. Bi-directional Interaction

In a mini-batch, the Siamese network takes as inputs scene images and corresponding cropped patches. Therefore, for a positive RoI, the two branches should have consistent responses to it. We believe that the consistency exists at two levels, including the feature-level and the prediction-level. For a positive RoI, the former means that in the feature space, it should be embedded closely while the latter means that the two branches output the same identity prediction.

**Feature-level interaction.** For the feature-level consistency, a feature-level interaction loss is defined as the cosine similarity between the features in two branches, computed as

$$L_{fi} = \frac{1}{N(\Omega)} \sum_{b \in \Omega} [1 - \cos(f_r(b), f_s(b))], \tag{1}$$

where $\Omega$ denotes the sets of positive RoIs sampled, $f_r(b)$ and $f_s(b)$ the features in the instance-aware and search branch, respectively.

**Prediction-level interaction.** For the prediction-level consistency, a prediction-level interaction loss is defined as the Kullback Leibler (KL) Divergence:

$$L_{pi} = D_{KL}(\tilde{p}_r(b) || \tilde{p}_s(b)) + D_{KL}(\tilde{p}_s(b) || \tilde{p}_r(b)), \tag{2}$$

where the KL divergence from $p_s(b)$ to $p_r(b)$ is computed as

$$D_{KL}(\tilde{p}_r(b) || \tilde{p}_s(b)) = \sum_{b \in \Omega} \sum_{c=1}^{C} \tilde{p}_r^c(b) \log \frac{\tilde{p}_r^c(b)}{\tilde{p}_s^c(b)}, \tag{3}$$

where $C$ is the length of the look-up table in OIM loss. The soften probability distribution is computed as:

$$\tilde{p}^i = \frac{\exp(p^i/T)}{\sum_{c=1}^{C} \exp p^c/T}, \tag{4}$$

where $T$ is the temperature. The $p^c$ denotes the probability of $b$ belonging to class $c$, calculated by the OIM loss.

Through the interaction losses $L_{fi}$ and $L_{pi}$, BINet achieves bi-directional interaction between two branches and learns more discriminative identity features.

The overall learning objective function is given as:

$$L = L_{det} + L_i + 0.5 * (L^s_{oim} + L^r_{oim}), \quad (5)$$

where $L_{det}$ stands for the detection loss used in Faster-RCNN [23].The loss $L_i$ is the sum of $L_{fi}$ and $L_{pi}$. The learning of identity features in the search and instance-aware branches is supervised by $L^s_{oim}$ and $L^r_{oim}$, respectively.

### 3.5. Discussion

The proposed BINet shares similar motivations with previous works [4, 37], because they all aim to alleviate the burden of the context information outside bounding boxes through a two-branch framework. However, BINet differs from them significantly from the following aspects:

(a) BINet is proposed to alleviate the negative influence of context on person Re-id rather than the classification task in the generic object detection. Compared with generic classification, person Re-id is a fine-grained task, which is more likely to be affected.

(b) In [4], the authors train an additional R-CNN branch besides the Faster R-CNN branch and merge the classification scores of both to improve the detection performance. Due to some arguments, the parameters of these two networks are not shared. Therefore, during inference, it cost more time to apply both branches. However, BINet is a Siamese Network, and only the major branch is applied during inference.

(c) In Deformable-v2 [37], the authors introduce a feature mimic loss to force the features of Deformable Faster R-CNN to be similar to R-CNN features extracted from cropped images. In the training procedure, the gradient between the two branches is uni-directional. Differently, our interaction is bi-directional and we apply a prediction-level constraint besides the feature-level constraint.

## 4. Experiments

### 4.1. Datasets

**CUHK-SYSU:** CUHK-SYSU [26] is a large scale person search dataset consisting of street snaps shot by hand-held cameras and snapshots collected from movies. It contains 18,184 scene images, 8,432 labeled identities and 96,143 annotated bounding boxes. Each labeled identity is assigned a class-id and appears in at least two different scene images from different viewpoints. The unlabeled identities are marked as unknown persons. The training set contains 11,206 scene images and 5,532 query persons, while the testing set includes 6,978 gallery images and 2,900 query persons. In the testing set, for each query person, there is a set of protocols with gallery size ranging from 50 to 4,000.

**PRW:** PRW dataset [36] contains 11,816 video frames extracted from one 10-hour video captured on a university campus. It contains 932 identities and 34,304 annotated bounding boxes. Similar to CUHK-SYSU, all proposals are divided into two groups, labeled identities and unlabeled identities. The training set includes 5,704 images and 482 different persons, while the testing set contains 6,112 images and 2,057 probe persons from 450 different identities. For each query person in the testing set, the search space is the whole gallery set.

### 4.2. Evaluation Protocol

We adopt the Cumulative Matching Characteristic (CMC) and the mean Averaged Precision (mAP) as performance metrics, which are the same as the previous works [21, 3]. The first metric is widely used in classification, where a matching is counted if there is at least one of the top-K predicted bounding boxes overlapping with the ground truth with an IoU larger or equal to 0.5. The second metric is widely used in object detection. We calculate an averaged precision (AP) by computing the area under the Precision-Recall curve for each query person and then average the APs across all the queries to obtain the mAP.

### 4.3. Implementation Details

We use PyTorch to implement our model, and run the experiments on the NVIDIA 1080Ti GPU. The ResNet-50 based BINet is initialized with an ImageNet [6] pre-trained model. For training the end-to-end model, we adopt SGD algorithm with the momentum set to 0.9, the weight decay to 0.0001, and the batch size to 2. For CUHK-SYSU, the scene images are resized to have at least 600 pixels on the short side and at most 1,000 pixels on the long side. The learning rate is initialized to 0.001, dropped to 0.0001 after 40K iterations, and kept unchanged until 50K iterations. For PRW, the images are resized to have at least 900 pixels on the short side and at most 1,500 pixels on the long side. The learning rate is initialized to 0.001, dropped to its 1/10 after 20K iterations, and kept unchanged until 30K iterations. The circular queue size is set to 5,000 and 500 for training CUHK-SYSU and PRW, respectively. Other details are the same as the previous work [26].

### 4.4. Ablation Study

In this subsection, we perform several analytic experiments on CUHK-SYSU and PRW to explore the contribution of each component in our proposed BINet, including the instance-aware branch and the bi-directional interaction losses.

**Effectiveness of BINet.** In Table 1, we show the effectiveness of two key components in our proposed BINet from

Table 1. Results of two key components on CUHK-SYSU and PRW. Legend: Detected: the proposals are detected by the multi-task framework during testing phase; Labeled: the model is tested with the ground truth bounding boxes.

| Dataset | CUHK-SYSU | | | | PRW | |
|---|---|---|---|---|---|---|
| Gallery Size | 100 | | 4000 | | 6112 | |
| Method | mAP(%) | top-1(%) | mAP(%) | top-1(%) | mAP(%) | top-1(%) |
| Detected | | | | | | |
| Baseline | 86.4 | 87.2 | 66.4 | 68.8 | 35.0 | 74.1 |
| + Instance-aware | 88.3 | 88.9 | 70.2 | 72.4 | 39.7 | 77.8 |
| + Interaction (BINet) | **90.0** | **90.7** | **74.6** | **77.2** | **45.3** | **81.7** |
| Labeled | | | | | | |
| Baseline | 87.4 | 87.9 | 67.6 | 70.0 | 37.0 | 76.3 |
| + Instance-aware | 89.2 | 89.8 | 71.2 | 73.7 | 41.6 | 79.6 |
| + Interaction (BINet) | **90.8** | **91.6** | **75.4** | **78.1** | **47.2** | **83.4** |

an overall view. As aforementioned, we modify the framework in OIM [26] and take it as our baseline. In the method named "+ Instance-aware", the framework is a Siamese network without interaction, taking as inputs scene images and cropped person patches from them. The results show that the "clean" data can benefit the person search performance significantly on both datasets. For example, on CUHK-SYSU dataset with 4000 gallery size setting, the Siamese network improves mAP by 3.8% and top-1 by 3.6 %. On PRW dataset, it improves mAP by 4.7% and top-1 by 3.7%. These results demonstrate our motivation that the model needs the guidance of the person patches to discriminate human appearance information from redundant context information.

In the method named "+ Interaction (BINet)", we introduce the interaction losses proposed in Sec 3.4 to the Siamese network. Through the losses, BINet can perform information interaction between two branches and achieve further improvements based on the Siamese network. Compared with the baseline, BINet achieves more than 8% improvement on mAP on both datasets (66.4% $\rightarrow$ 74.6%, 35.0% $\rightarrow$ 45.3% ). The results of using ground truth bounding boxes are consistent with that of using detected proposals. These results suggest that with the guidance of the cropped patches and the information interaction between two branches, BINet can focus on the persons rather than context, so that it can learn discriminative features of instances from scene images.

**Ablation study on different interaction settings.** In BINet, we introduce two interaction losses to achieve bi-directional interaction between branches in the feature-level and prediction-level, and the feature-level interaction loss is computed on the 2048-d features before the embedding layer. We compare results with different interaction settings in training BINet:

- BINet-0: we remove both interaction losses.
- BINet-1: the interaction is uni-directional, i.e., from the instance-aware branch to the search branch.
- BINet-2: the interaction is uni-directional, i.e., from the search branch to the instance-aware branch.
- BINet-3: only the feature-level interaction is kept.
- BINet-4: only the prediction-level interaction is kept.

Table 2. Results of different interaction settings on CUHK-SYSU and PRW.

| Dataset | CUHK-SYSU | | PRW | |
|---|---|---|---|---|
| Method | mAP(%) | top-1(%) | mAP(%) | top-1(%) |
| BINet | **74.6** | **77.2** | **45.3** | **81.7** |
| BINet-0 | 70.2 | 72.4 | 39.7 | 77.8 |
| BINet-1 | 73.5 | 76.1 | 43.3 | 80.3 |
| BINet-2 | 69.3 | 71.8 | 36.0 | 75.3 |
| BINet-3 | 73.3 | 76.1 | 42.3 | 78.8 |
| BINet-4 | 72.2 | 74.7 | 42.9 | 80.9 |

Results are shown in Table 2, from which we make the following observations: **(a)** The experiments of BINet, BINet-0, BINet-3 and BINet-4 explore the effectiveness of individual interaction losses. We find that either feature-level or prediction-level interaction can improve the performance. With both of them, the model achieves the best performance on both datasets. This is because these two interaction losses convey different information at different levels. **(b)** In BINet-1, the uni-directional interaction from the instance-aware branch to the search branch gains a 3.3% increase in mAP on CUHK-SYSU while the opposite interaction harms the performance in BINet-2. These results also demonstrate that the context information keeps the model from learning discriminative features. The best results achieved by BINet indicate that bi-directional interaction helps the model learn the most discriminative features.

**Influence of different temperature.** We evaluated the impact of the temperature setting in Eq. 4 in the range from 0.3 to 3.0. The results in Table 3 show that on CUHK-SYSU, $T = 0.3$ achieves the best performance while on PRW, the best choice of $T$ is 1.0. This is because the number of identities in CUHK-SUSY is 10 times more than that in PRW.

Table 3. Results of different temperature settings on CUHK-SYSU and PRW.

| Dataset | CUHK-SYSU | | PRW | |
|---|---|---|---|---|
| T | mAP(%) | top-1(%) | mAP(%) | top-1(%) |
| 0.3 | **74.7** | **77.4** | 43.9 | 80.9 |
| 0.5 | 74.6 | 77.2 | 45.0 | 81.4 |
| 1.0 | 74.6 | 77.2 | **45.3** | **81.7** |
| 3.0 | 73.7 | 76.4 | 44.5 | **81.7** |

Therefore, on CUHK-SYSU, the original probability distribution calculated by OIM loss is so soft that the prediction-level interaction loss needs a lower temperature.

### 4.5. Comparison with the State-of-the-arts

Table 4. Comparison of performance on CUHK-SYSU with 100 gallery size setting.

| Method | mAP(%) | top-1(%) |
|---|---|---|
| OIM [26] | 75.5 | 78.7 |
| IAN [24] | 76.3 | 80.1 |
| NPSM [19] | 77.9 | 81.2 |
| RCAA [2] | 79.3 | 81.3 |
| $CNN_v$ + MGTS [3] | 83.0 | 83.7 |
| CNN + CLSA [15] | 87.2 | 88.5 |
| CNN + Refinement [11] | **93.0** | **94.2** |
| Context [29] | 84.1 | 86.5 |
| QEEPS [21] | 88.9 | 89.1 |
| OIM (ours) | 86.4 | 87.2 |
| BINet (ours) | **90.0** | **90.7** |

**Evaluation on CUHK-SYSU.** Table 4 shows the person search results on CUHK-SYSU with a gallery size of 100. The notations "$CNN_v$" and "CNN" denote the Faster R-CNN detector based on VGGNet and ResNet-50, respectively. Our modified OIM outperforms most of the previous methods, including the two-stage method "$CNN_v$ + MGTS". This demonstrates the effectiveness of training detection and person Re-id jointly. Compared with modified OIM, our proposed BINet obtains the performance gain of 3.6%/3.5% in terms of mAP/top-1 and introduces no additional computation during inference, which demonstrates the importance of removing the context information outside bounding boxes in the training procedure of end-to-end methods. Compared with the prior best query-guided method QEEPS, our BINet also outperforms it. Moreover, our method is not query-guided, which makes it more efficient during inference. We observe that the results in [11] are better than ours. [11] applies two ResNet50-based models to handle the pedestrian detection and person Reid respectively and adopt many tricks [20] in the baseline method. However, in our baseline method, we need to solve the detection and Re-id in a single multi-task networks, so we cannot apply those tricks due to the detection part. The mAP performance of their baseline is 5.8% better than our baseline on CUHK-SYSU (92.2% vs. 86.4%). Therefore, it is acceptable that the results in [11] are better than ours. Compared to [11], our method can not only simplify the training procedure but also save parameters.

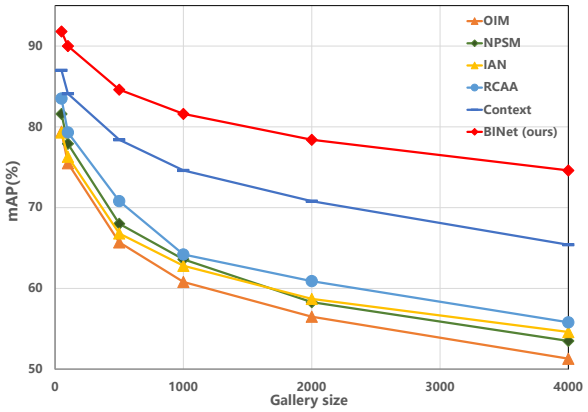


Figure 4. Evaluation on CUHK-SYSU with different gallery sizes.

To evaluate the scalability of our method, we compare with other end-to-end methods under different gallery sizes in the range from 50 to 4,000. As shown in Figure 4, all the methods degrade the performance with the gallery size increasing. This is because more distractors are involved when the gallery becomes larger. We can observe that our BINet still outperforms other methods under all gallery sizes. Moreover, when increasing the gallery size from 50 to 4,000, the mAP performance of modified OIM drops from 89.0% to 66.4% while our BINet drops from 91.8% to 74.6%. This verifies the robustness of our method.

Table 5. Comparison of performance on PRW.

| Method | mAP(%) | top-1(%) |
|---|---|---|
| OIM [26] | 21.3 | 49.9 |
| IAN [24] | 23.0 | 61.9 |
| NPSM [19] | 24.2 | 53.1 |
| $CNN_v$ + MGTS [3] | 32.6 | 72.1 |
| CNN + CLSA [15] | 38.7 | 65.0 |
| CNN + Refinement [11] | 42.9 | 70.2 |
| Context [29] | 33.4 | 73.6 |
| QEEPS [21] | 37.1 | 76.7 |
| OIM (ours) | 35.0 | 74.1 |
| BINet (ours) | **45.3** | **81.7** |

**Evaluation on PRW.** We further evaluate the BINet on the PRW dataset. Overall, we observed similar performance comparisons with state-of-the-art methods as on CUHK-SYSU. Specifically, our BINet still achieves the best person search performance among these methods, which surpasses the prior best model QEEPS by 8.2% and 5.0% in mAP and top-1, respectively. This consistently demonstrates the superiority of our proposed method.

### 4.6. Further Analysis

In this section, we further analyze the influence of the cropped patches and the interaction on representation learning. To exclude the influence of different detections, the models are evaluated with the ground truth bounding boxes.

**Analysis of the cropped patches.** In the procedure of training BINet, for the positive RoIs obtained from the RPN, the corresponding patches are cropped and resized to $224 \times 96$ before being fed into the instance-aware branch. For the feature learning part of our proposed framework, the scale variation of persons is augmented. Therefore, the improvements brought by the cropped patches may come from two aspects: **scale augmentation** and **removing context information**. To figure out the influence of each part, we remove the scale augmentation and report the results in Table 6. In the method Re-id-O and BINet-O, we keep the original sizes of the cropped patches unchanged and apply an RoI-Align layer after the first part in the instance-aware branch. In this way, for each RoI, the scale is the same in both branches so that the scale augmentation is removed at the training stage. From the performance of Re-id-O and BINet-O, we observe that our method still improves the search performance significantly. For example, BINet-O improves mAP by 9.9% and top-1 by 7.6% on PRW. When comparing the methods using resized patches (Re-id-R and BINet-R), we find that almost no performance gains on PRW are observed. On CUHK-SYSU, the gains are also marginal compared to improvements brought by removing context information. Based on the above observations, we can conclude that in the procedure of training BINet, the data without the context information outside the bounding boxes plays a more important role than scale augmentation.

Table 6. The influence of different scale settings on performance on CUHK-SYSU with gallery size 4000 and PRW. "Re-id" stands for the Siamese network with no interactions.

| Dataset | CUHK-SYSU | | PRW | |
|---|---|---|---|---|
| Method | mAP(%) | top-1(%) | mAP(%) | top-1(%) |
| Baseline | 67.6 | 70.0 | 37.0 | 76.3 |
| Re-id-O | $70.9_{\uparrow 3.3}$ | $73.3_{\uparrow 3.3}$ | $41.8_{\uparrow 4.8}$ | $79.8_{\uparrow 3.5}$ |
| Re-id-R | $71.2_{\uparrow 3.6}$ | $73.7_{\uparrow 3.7}$ | $41.6_{\uparrow 4.6}$ | $79.6_{\uparrow 3.3}$ |
| BINet-O | $73.8_{\uparrow 6.2}$ | $76.4_{\uparrow 6.4}$ | $46.9_{\uparrow 9.9}$ | $83.9_{\uparrow 7.6}$ |
| BINet-R | $75.4_{\uparrow 7.8}$ | $78.1_{\uparrow 8.1}$ | $47.2_{\uparrow 10.2}$ | $83.4_{\uparrow 7.1}$ |

**Bi-direction vs. Uni-direction.** In order to better understand the effectiveness brought by the bi-directional interaction, we evaluate the instance-aware branch with the resized person patches cropped from the original scene images and report the mAP. As shown in Table 7, the comparisons are consistent with those in Table 2. With the bi-directional interaction, both branches achieve the best performance. These results show that the model with bi-directional interaction learns the most discriminative features.

Table 7. Experimental comparisons for different interaction directions. The arrows stand for the directions of the gradients between two branches. The legend "S" and "R" denotes the performance of the search and instance-aware branch, respectively.

| Dataset | CUHK-SYSU | | PRW | |
|---|---|---|---|---|
| Interaction | S | R | S | R |
| no | 71.2 | 72.3 | 41.6 | 45.7 |
| S$\longrightarrow$R | 70.4 | 72.0 | 38.3 | 42.3 |
| R$\longrightarrow$S | 74.2 | 74.6 | 45.1 | 47.1 |
| S$\longleftrightarrow$R | 75.4 | 75.6 | 47.2 | 49.9 |

**Runtime comparison.** In Table 8, we report the time taken by QEEPS [21], original OIM [26], modified OIM*, and BINet to process a gallery image. Since we reduce the output size of the RoI-Align layer and the residual blocks in the second part, the modified OIM* (baseline) and is more efficient, i.e., it is more than 3 times faster than QEEPS. Moreover, our BINet is not query-guided, so all the queries can share the proposals generated from the gallery scene images, which makes it more practical than QEEPS in real-world applications.

Table 8. Runtime comparison of BINet with others for image size $900 \times 1500$. * stands for our modified version.

| Method | GPU | Time($sec$) |
|---|---|---|
| QEEPS [21] | P6000 | 0.30 |
| OIM [26] | 1080Ti | 0.17 |
| OIM*(ours) | 1080Ti | 0.08 |
| BINet(ours) | 1080Ti | 0.08 |

## 5. Conclusion

In this paper, we propose a Siamese network, named Bi-directional Interaction Network, which takes as inputs scene images and cropped person patches. With the guidance of cropped patches, BINet can focus on the persons in the scenes. We also design interaction losses to achieve bi-directional information interaction between branches. Extensive experiments demonstrate that our method can significantly improve the performance without additional computation during inference.

## Acknowledgements

This work was supported in part by the National Key R&D Program of China (No.2018YFB1402600), the National Natural Science Foundation of China (No.61836014, No.61761146004, No.61773375, No.61602481), the Key R&D Program of Shandong Province (Major Scientific and Technological Innovation Project) (NO.2019JZZY010119), and CAS-AIR.

## References


[1] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, pages 2654–2662, 2014. 3

[2] Xiaojun Chang, Po-Yao Huang, Yi-Dong Shen, Xiaodan Liang, Yi Yang, and Alexander G Hauptmann. Rcaa: Relational context-aware agents for person search. In *Proceedings of the European Conference on Computer Vision*, pages 84–100, 2018. 2, 3, 7

[3] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *Proceedings of the European Conference on Computer Vision*, pages 734–750, 2018. 2, 5, 7

[4] Bowen Cheng, Yunchao Wei, Honghui Shi, Rogerio Feris, Jinjun Xiong, and Thomas Huang. Revisiting rcnn: On awakening the classification power of faster rcnn. In *Proceedings of the European Conference on Computer Vision*, pages 453–468, 2018. 2, 5

[5] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[7] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014. 2

[8] Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge Belongie. Integral channel features. 2009. 2

[9] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 2

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 2

[11] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-id driven localization refinement for person search. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9814–9823, 2019. 2, 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 3

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *stat*, 1050:9, 2015. 3

[14] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, 2012. 2

[15] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In *Proceedings of the European Conference on Computer Vision*, pages 553–569, 2018. 2, 7

[16] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6356–6364, 2017. 3

[17] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 2

[18] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan 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*, pages 2197–2206, 2015. 2

[19] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 493–501, 2017. 2, 7

[20] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 7

[21] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 3, 5, 7, 8

[22] Woonhyun Nam, Piotr Dollár, and Joon Hee Han. Local decorrelation for improved pedestrian detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2014. 2

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 2, 5

[24] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. Ian: the individual aggregation network for person search. *Pattern Recognition*, 87:332–340, 2019. 1, 2, 7

[25] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016. 2

[26] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017. 1, 2, 3, 5, 6, 7, 8

[27] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional

lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, pages 802–810, 2015. 2

[28] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 937–940. ACM, 2014. 1, 2

[29] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 7

[30] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z Li. Convolutional channel features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 82–90, 2015. 2

[31] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, pages 34–39. IEEE, 2014. 2

[32] Shanshan Zhang, Christian Bauckhage, and Armin B Cremers. Informed haar-like features improve pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 947–954, 2014. 2

[33] Shanshan Zhang, Rodrigo Benenson, Bernt Schiele, et al. Filtered channel features for pedestrian detection. In *CVPR*, volume 1, page 4, 2015. 2

[34] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013. 2

[35] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 2

[36] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017. 2, 5

[37] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 3, 5