Instance Guided Proposal 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

Person detection networks have been widely used in person search. These detectors discriminate persons from the background and generate proposals of all the persons from a gallery of scene images for each query. However, such a large number of proposals have a negative influence on the following identity matching process because many distractors are involved. In this paper, we propose a new detection network for person search, named Instance Guided Proposal Network (IGPN), which can learn the similarity between query persons and proposals. Thus, we can decrease proposals according to the similarity scores. To incorporate information of the query into the detection network, we introduce the Siamese region proposal network to Faster-RCNN and we propose improved cross-correlation layers to alleviate the imbalance of parameters distribution. Furthermore, we design a local relation block and a global relation branch to leverage the proposal-proposal relations and query-scene relations, respectively. Extensive experiments show that our method improves the person search performance through decreasing proposals and achieves competitive performance on two large person search benchmark datasets, CUHK-SYSU and PRW.

1. Introduction

Person search aims to localize a target matching the query person in a gallery of whole unconstrained scene images. It is extended from person re-identification (Re-id) whose goal is to match the query in a gallery of manually cropped or carefully filtered auto-detected person patches. It has many applications in the real world, 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*.

*Corresponding Author.

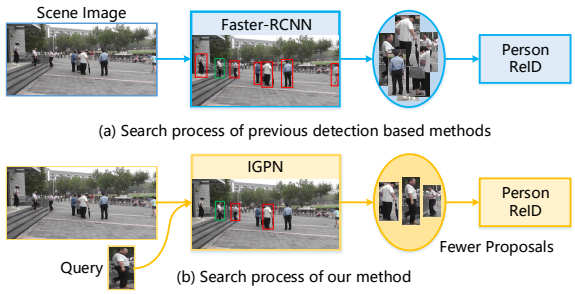

Figure 1. Comparison with previous detection based methods. (a) Previous detection based methods first crop proposals of all the persons in gallery scene images and feed them into a person Re-id model to calculate the similarity scores for each proposal. (b) Our proposed IGPN leverages the appearance information of the queries and learn the similarity scores between proposals and queries. Thus, we can decrease the proposals fed into the person Re-id model by keeping the proposals with high similarity scores.

Following [26], some works [24, 29, 23, 14, 11, 3, 2] have been proposed to address this problem. According to the ways of obtaining person patches from scene images, existing methods can be grouped into two categories: detection based methods [24, 23, 11, 3] and search based methods [14, 2]. Detection based methods are composed of two parts: person detection and person Re-id networks, which are trained jointly or separately. Generally, for each query, the person detection network is used to detect candidate persons within scene images, and then the person Re-id network is applied to calculate the similarity between all pairs of the query and the candidates. However, this pipeline has a drawback. In the first stage, the detection network generates proposals of all the persons. In the community of person Re-id, it is commonly believed that with the gallery size increasing, person Re-id becomes more challenging because more distractors are involved. Thus, such a large number of proposals will have a negative influence on the following identity matching process.

Instead of cropping all the persons, search based meth-

ods recursively shrink the search region to more accurately locate the target person in the scene with the guidance of information of the query. Although there are much fewer proposals obtained from scenes in this pipeline, search based methods have other drawbacks. Firstly, these methods only locate one person for each scene image. If the target person has another person with similar appearance around, the model may locate the wrong person, which leads to misdetections. Secondly, to locate the target person accurately, it is inefficient to run the model for many times to shrink the initial search region (the whole scene).

Therefore, for person search, we need a new method of generating proposals from scenes, which shares the merits of two kinds of methods. More specifically, it can not only leverage the query information to decrease proposals like search based methods, but also preserve multiple proposals for a scene like detection based methods if necessary (*i.e.*, there are some distracting people around the target). In addition, it should also be more efficient than search based methods.

Motivated by the above observations, we propose a new person detection network for person search, named Instance Guided Proposal Network (IGPN). Compared with other person detection networks in the existing methods (Figure 1), our IGPN introduces the appearance information of the query person and outputs similarity scores for each proposal. Given a gallery of scene images, we can keep the candidates similar to the target person by ranking the corresponding similarity scores to decrease proposals. Compared with the search based methods, our method reduces the misdetection rate and is more efficient, because the model only needs to be run once to locate the target person in a scene.

To incorporate information of the query into the detection network, we introduce the Siamese region proposal network (Siamese-RPN) to Faster-RCNN [18]. Different from the vanilla Siamese-RPN [13], we propose *improved cross-correlation layers* (ICCL) to alleviate the imbalance of parameters distribution while maintaining the performance. In [9, 19, 15, 2], it is found that relationships can benefit various tasks. Inspired by these works, our IGPN leverages relations through two ways. The first is a *local relation block* proposed to model the proposal-proposal relations. The second is a *global relation branch* to characterize the query-scene relations.

To summarize, we make the following contributions to person search:

- We propose a new person detection network, named Instance Guided Proposal Network (IGPN), which integrates the query person information into the detection network to learn the similarity between person proposals and the target person.
- We propose ICCL to alleviate the imbalance of parameters distribution in the vanilla Siamese-RPN without loss of performance.
- We design a local relation block and a global relation branch to leverage the proposal-proposal and query-scene relations, respectively.

2. Related Work

Person search. Person search 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 [24] and PRW [29]. These methods can be grouped into two categories, detection based [24, 23, 29, 11, 3] methods and search based methods [14, 2]. Detection based methods are composed of two aspects: pedestrian detection and person re-identification. In [24], [23] and [16], they both develop an end-to-end person search framework based on Faster R-CNN [18] to jointly handle the two aspects. Yan *et al.* [27] built a graph on the top of an end-to-end network to learn context information. Lan *et al.* [11] 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. Different from the detection based methods, search based methods 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.* [14] propose Conv-LSTM [25] 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. However, existing methods do not pay much attention to the problem that a large number of proposals obtained from scene images have a negative influence on the overall search performance.

Person detection. For the detection networks used in person search, [24] and [29] compare the effect of different person detectors (*e.g.* DPM [6], ACF [5], CCF [28], and LDCF [17]) on the overall search performance. More recently, [3] and [11] apply Faster-RCNN [18] to detect persons in scenes. However, all the detectors above can only discriminate persons from the background and leads to a large number of proposals that suppress the improvement of the search performance.

Siamese networks. Recently, Siamese networks have drawn significant attention in visual tracking due to their balanced accuracy and efficiency. These trackers [1, 21, 30, 13, 12] formulate visual tracking as a cross-correlation problem. To incorporate the appearance information of the query person into the detection network, we combine Faster R-CNN with Siamese-RPN and propose an Instance Guided

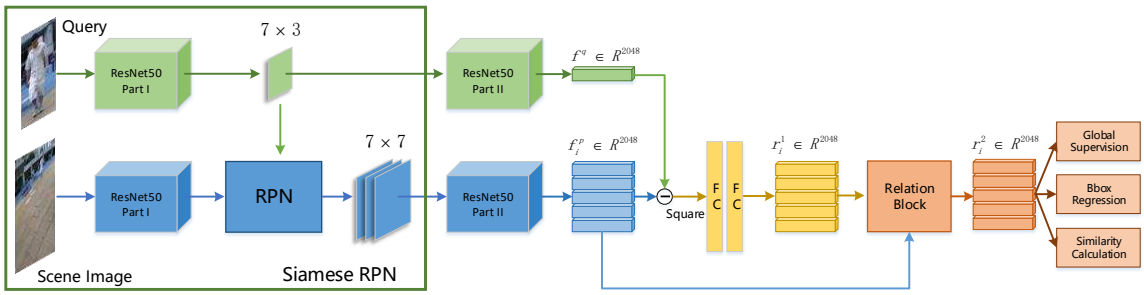

Figure 2. Illustration of our proposed IGPN. "ResNet50 PartI" corresponds to the $conv1$ to $conv4$ of ResNet50 [8] while "ResNet50 PartII" represents the $conv5$ of ResNet50. And the "ResNet50 PartI" and "ResNet50 PartII" in two branches share parameters. For brevity, we do not show the channel-dimension of the feature maps, the 2×2 max-pooling layer before the 7×3 feature map, the ROI pooling layer, the similarity calculation and bounding box regression loss in the RPN module in the figure above.

Proposal Network (IGPN) for person search.

3. Instance Guided Proposal Network

3.1. Overview

The overall architecture of IGPN is shown in Figure 2. We choose ResNet50 [8] as the backbone. The proposed IGPN mainly consists of an improved Siamese region proposal network (Siamese-RPN) and a local relation block which leverages appearance information of the query and relationships between pairs of proposals in the same scene, respectively. Moreover, it can exploit the global relationship between the query and scenes through the global relation branch. It takes a pair of a query person patch and a scene image as input and outputs bounding boxes along with similarity scores.

When taken as a person detection network, IGPN works with a separately trained person Re-id network. Given a query person and a set of gallery scene images, we first obtain many proposals through IGPN. Then we remove the proposals with low similarity scores. Only the remaining ones are fed into the Re-id network. Therefore, compared with the detection networks in methods [3, 11], our IGPN can decrease proposals to benefit the overall search performance. Moreover, our method can preserve several candidates in a scene if there are some distracting factors, which avoids the mis-detections caused by the search strategy in search based methods [14, 2].

3.2. Siamese-RPN

As shown in Figure 2, Siamese-RPN is composed of two parts: a Siamese feature extractor and a region proposal network (RPN). Taking a pair of a scene image and a query person patch, Siamese-RPN generates a set of refined anchors with similarity scores. Different from the vanilla Siamese-RPN in [13, 30], we improve the cross-correlation layers to alleviate the imbalance of parameters distribution.

3.2.1 Feature Extraction

Firstly, we will give a brief introduction to the feature extractor. The Siamese feature extractor consists of two branches, one branch for learning the feature representation of the query person patch and the other for the scene image. The two branches share parameters in CNN so that the two images are encoded in the same semantic embedding space. For convenience, we denote z and x as the feature maps of the query patch and scene image, respectively.

3.2.2 Improved Cross-correlation Layers

In the RPN module, there are two branches, one for the first query-anchor similarity calculation, the other for the first bounding box regression. Similar to the conventional Faster-RCNN, two convolution layers are applied to adjust x to $\phi_{sim}(x)$ and $\phi_{reg}(x)$. Two correlation kernels $\varphi_{sim}(z)$ and $\varphi_{reg}(z)$ are obtained by using another two convolution layers. Then the similarity scores $A_{sim} \in R^{H \times W \times 2k}$ and regression offsets $A_{reg} \in R^{H \times W \times 4k}$ can be computed as

$$\begin{aligned} A_{sim} &= \phi_{sim}(x) \star \varphi_{sim}(z) \\ A_{reg} &= \phi_{reg}(x) \star \varphi_{reg}(z) \end{aligned} \tag{1}$$

where k is the number of anchors and $\star$ denotes the correlation operation. In [13], the convolution operations $\varphi_{sim}(\cdot)$ and $\varphi_{reg}(\cdot)$ directly increases the channels of z to obtain the correlation kernels $\varphi_{sim}(z)$ and $\varphi_{reg}(z)$, respectively. However, such operations lead to severe imbalance of parameter distribution. For example, as shown in Figure 3, a 3×3 convolution kernel in $\varphi(\cdot)_{sim}$ contains $3 \times 3 \times c \times (c' \times 2k)$ parameters, where c and c' are the numbers of channels of z and $\phi(x)$. In our methods, k, c and c' are set to 9, 1024 and 512, respectively, which causes serious imbalance of parameter distribution (*i.e.*, the RPN module contains 254M parameters while the whole ResNet50 contains 25M parameters). To address this issue, instead

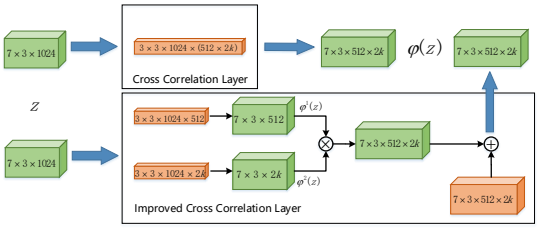

Figure 3. The architecture of the vanilla and improved cross-correlation layers. Green cubes denote the feature maps and orange ones denote the parameters to learn. "⊗" denotes outer product and "⊕" denotes element-wise sum.

of using the up-channel operation directly, we compute the correlation kernels as

$$\begin{aligned}\varphi_{sim}(z) &= \varphi^1_{sim}(z) \otimes \varphi^2_{sim}(z) + B_{sim} \\ \varphi_{reg}(z) &= \varphi^1_{reg}(z) \otimes \varphi^2_{reg}(z) + B_{reg}\end{aligned} \tag{2}$$

where $\otimes$ denotes outer product. The difference between the two types of cross-correlation layers is shown in Figure 3. For brevity, we only explain the operation in the similarity calculation branch and remove the subscripts. In ICCL, we divide the correlation kernels $\varphi(z)$ into two parts. The first part is the outer product of two feature maps $\varphi^1(z) \in R^{h\times w\times c'}$ and $\varphi^2(z) \in R^{h\times w\times 2k}$, which changes according to different query persons. The second part is $B \in R^{h\times w\times c'\times 2k}$, which is shared by all the query persons and fixed in the inference stage. Therefore, as shown in Figure 3, in the ICCL, the model only needs to learn $\varphi^1(\cdot) \in R^{3\times 3\times c\times c'}$, $\varphi^2(\cdot) \in R^{3\times 3\times c\times 2k}$ and $B \in R^{h\times w\times c'\times 2k}$. As shown in Eq. 2, the first item $\varphi^1_{sim}(z) \otimes \varphi^2_{sim}(z)$ changes according to different query persons so that it can learn the specific information from each query (e.g. colors of clothes). Meanwhile, the second item B_{sim} is shared by all query persons to learn common information from them (e.g. information of human body). The novel cross-correlation layers only contain 10M parameters but can maintain the performance simultaneously.

3.3. Multiple Relation Modeling

In IGPN, we design two parts to leverage relations. The first is a local relation block proposed to model the *proposal-proposal* relations. The second is a global relation branch to characterize the *query-scene* relations. For the Local Relation Block, the contribution is that we redesign it with a Siamese network in our paper. For the Global Relation Branch, the key contribution is that we introduce an auxiliary task detailed in the following subsection, which is a novel way to exploit the relationship between the query and the scene in the person search area.

3.3.1 Local Relation Block

As shown in Figure 2, after the global average pooling layer in "ResNet Part II", we can obtain a feature f^q for the query and N features $\{f^p_i\}^N_{i=1}$ for N proposals. Each f^p_i is subtracted by f^q and processed by element-wise square operation and two fully connected layers to obtain the relation features r^1_i. Then features of proposals and relation features are fed into a local relation block to exploit the relation information between the proposals in the same scene image. The local relation block is similar to non-local block [22]. However, we redesign it with a Siamese network, where the inputs to the block are two different sets of features $\{r^1_i\}$ and $\{f^p_i\}$, which is significantly different from existing methods. Moreover, we replace the pixel-wise relationships in the standard non-local block [22] with relationships between the features of proposals. More specifically, the relationships are computed in an embedded Gaussian version:

$$R(f^p_i, f^p_j) = e^{\mu(f^p_i)^T \nu(f^p_j)} \tag{3}$$

Here $\mu(f^p_i) = W_\mu f^p_i$ and $\nu(f^p_j) = W_\nu f^p_j$ are two embeddings. i and j are the indexes of the proposals. Then the relation features r^1_i are refined with the relationships to obtain the new relation features r^2_i:

$$r^2_i = W_r \tilde{r}^2_i + r^1_i \tag{4}$$

where W_r is the fully connected layer, "+" denotes the residual connection and $\tilde{r}^2_i$ is calculated as:

$$\tilde{r}^2_i = \frac{1}{Z}\sum_{j=1}^{N} R(f^p_i, f^p_j) \cdot g(r^1_j) \tag{5}$$

where Z denotes the normalization and $g(\cdot)$ is also an embedding like $\mu(\cdot)$ and $\nu(\cdot)$. By this way, the relation features r^2_i encode not only the query-proposal relationships but also the relationships between proposals. The rich relation information will help IGPN calculate more accurate similarity scores, which has been proved in the experiments section. The relation features are then used for the second similarity calculation and bounding box regression for each proposal.

3.3.2 Global Relation Branch

Here we describe the global supervision branch in detail. This branch is designed with the intuition that global information from the whole scene may benefit the person search task. For example, if the model determines that there is no target in the scene, the model can give the proposals from this scene low similarity scores. Thus, besides the similarity calculation and bounding box regression tasks, we introduce an auxiliary binary classification task of determining whether there is a query person in the scene, so that the

model can learn the global relationship between the query person and the scene image. From a multi-task learning view, one task can benefit from another task, like Mask-RCNN [7]. Therefore, the auxiliary task can benefit the quality of the proposals generated by IGPN.

Specifically, the global feature $r^g \in R^{2048}$ is taken from an average of the relation features $\{r_i^2\}_{i=1}^N$. After one fully connected layer, r^g becomes a 2-dimension vector $\hat{y}$ and binary classification (search loss) is optimized on the ground-truth labels y:

$$L_{search} = -\sum_{i=1}^{2} y_i \left(\hat{y}_i - \log \sum_{j=1}^{2} \exp \hat{y}_j \right) \qquad (6)$$

3.4. Decreasing Proposals for Person Re-id

Compared with the person detection networks used in [24, 3], our IGPN can calculate the similarity scores between the query and proposals, and preserves the candidates similar to the query. Fewer proposals can benefit the identity matching process so that the overall person search performance is improved. We analyze the influence of the number of proposals on the final performance and show the effectiveness of our IGPN under different gallery size settings in the experiments section.

3.5. Model Training

Here we detail the training process of the proposed IGPN. Firstly, we introduce the training samples used in our method. Different from the person detection networks in [11, 3], our IGPN takes pairs of query person patches and scene images as inputs. Training pairs are obtained from large-scale person search datasets, CUHK-SYSU [24] and PRW [29]. There are three different types of training pairs: cross-scene positive pairs, self-scene positive pairs and negative pairs. For cross-scene positive pairs, we choose two different scene images containing the same person and take the cropped person patch from one scene as the query and the other as the search image. Likewise, a negative pair contains a person patch as the query and a scene image which does not contain the query. For self-scene positive pairs, we take one scene image as the search image and the person in it as the query. The positive pairs can make the model learn the appearance change of the same person and the negative pairs can help to suppress the similarity scores on distractors.

In the RPN, positive samples are defined as the anchors which have $IoU > 0.6$ with their corresponding ground truth. Negative ones are defined as the anchors with $IoU < 0.3$. We also limit at most 16 positive samples and totally 64 samples from one search image. After the RPN, we keep 32 proposals and then feed them into the RoI pooling layer.

Furthermore, the aspect ratios of the annotated bounding boxes mostly range from 1.0 to 0.25. Therefore, for anchors, we use 3 aspect ratios of $1:1$, $1:2$ and $1:3$, and 3 scales with box areas of 64^2, 128^2 and 256^2.

The model is trained end-to-end using the following loss functions:

$$L = L_{sim}^1 + L_{reg}^1 + L_{sim}^2 + L_{reg}^2 + L_{search} \qquad (7)$$

Here L_{sim}^1 and L_{sim}^2 are the loss functions for the first similarity calculation in the RPN module and the second after the relation block, respectively. Likewise, L_{reg}^1 and L_{reg}^2 are the bounding box regression loss functions. We refer readers to [13, 18] for more details. L_{search} is the search loss defined in Eq. 6.

3.6. Distinctions with QEEPS

Recently, Munjal *et al.* [16] have proposed a query-guided end-to-end method for the person search task (QEEPS). They build a QRPN based on Squeeze-Excitation module [10] with a similar motivation to us. They solve the person search in an end-to-end manner while we aim to solve the task through a detection + person Re-id pipeline. For end-to-end methods, although they can optimize jointly the detection and Re-id parts, yet it is hard to incorporate more fine-grained information into the networks, e.g. part information, due to the detection parts. However, in our two-stage method which is more flexible, we can exploit part information by using part-based Re-id methods, like PCB model [20].

4. Experiments

4.1. Datasets

CUHK-SYSU: CUHK-SYSU [24] 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 identities 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 testing set, for each query person, there are a set of protocols with gallery size ranging from 50 to 4,000.

PRW: PRW dataset [29] 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 also used in the previous works. The first metric is widely used in person Re-id, 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 ResNet50 based IGPN is initialized with an ImageNet [4] pretrained model. It is trained using SGD with a batch size of 16. 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 query person patches are first padded to have an aspect ratio of $3:7$ and then resized to 224×96 pixels to keep the raw aspect ratios of the bounding boxes. We apply an RoI-Pooling layer on the "Res50 part1" to pool a $7 \times 7 \times 1024$ region from the stem feature maps for each proposal. The stride of "Res50 part2" is set to 1. We train the model for 25 epochs with base learning rate initialized at 10^{-3} and decayed to 10^{-4} after 16 epochs. Moreover, all the Batch Normalization layers are frozen during training.

For training the person Re-id network, we use both annotated and detected boxes as [11]. The training images are augmented with horizontal flip and normalization. We set the batch size to 64 and the epoch to 60. The initial learning rate is set to 0.1 and decayed to 0.01 after 40 epochs. All person bounding boxes are resized to 384×128 pixels. The tensor obtained from the $conv5$ layer is divided into 6 pieces. We refer readers to [20] for more details.

4.4. Ablation Study

In this section, we first conduct several experiments to analyze the effect of each component in our proposed IGPN architecture, including the ICCL, the local relation block and the global relation branch. Then we study the influence of proposal and gallery size settings. Since the proposed IGPN outputs bounding boxes along with similarity scores, it is also evaluated with top-K and mAP metrics.

Effectiveness of ICCL. As aforementioned, the key to the Siamese-RPN is the cross-correlation layer, which can introduce the information of the query to the detection network. To demonstrate the effectiveness of our proposed improved cross-correlation layers (ICCL), we replace them with the vanilla ones (VCCL) in [13]. This setting leads to serious imbalance of parameter distribution (*i.e.*, the RPN module contains 254M parameters while the whole ResNet50 contains 25M parameters), which makes it hard to train the model. The results show that our proposed ICCL can address this problem because it contains only about 10M parameters but can maintain the performance simultaneously. we have also validated the importance of $B_{sim/reg}$. Without $B_{sim/reg}$, the ICCL performs worse (84.0% in mAP, 83.9% in top-1).

Table 1. The comparison of ICCL and VCCL

Method	parameters	mAP(%)	top1(%)
VCCL	254M	84.2	84.1
ICCL(proposed)	**10M**	**84.5**	**84.1**

Effectiveness of Relation. In IGPN, we design two parts to leverage relations. The first is a local relation block proposed to model the proposal-proposal relations. The second is a global relation branch to characterize the query-scene relations. In the method named "w/o L", we remove the local relation block and the relation features $\{r_i^1\}_{i=1}^N$ are directly divided into several branches for the following tasks. This setting causes that the model cannot make use of the relationships between the proposals in the same scene, which leads to the decrease of 3.7% (84.5-80.8) in mAP and 4.9% (84.1-79.2) in top-1. In the method named "w/o G", we remove the auxiliary task that determines whether there is a query in a scene by removing the search loss L_{search} so that the model cannot learn the global relationships between queries and scenes. This setting causes the decrease of 3.2% (84.5-81.3) in mAP and 3.2% (84.1-80.9) in top-1, which means the global relation is important. Furthermore, we remove both the local relation block and global relation branch in the method named "w/o L&G" and observe the more decrease of the performance than removing one of them. This indicates that the two different types of relationships are not able to be replaced with each other and they both play a vital role in the proposed model.

Influence of proposal and gallery settings. In this subsection, we take our proposed IGPN as a person detection network and combine it with a person Re-id network. Compared with the detection networks in the existing meth-

Table 2. Effect of leveraging two kinds of relation on CUHK-SYSU dataset with 100 gallery size setting. Legend: L: Local relation block, G: Global relation branch, "full" means that we keep both kinds of relations.

	L	G	mAP(%)	top-1(%)
full	✓	✓	**84.5**	**84.1**
w/o L		✓	80.8	79.2
w/o G	✓		81.3	80.9
w/o L&G			79.2	77.7

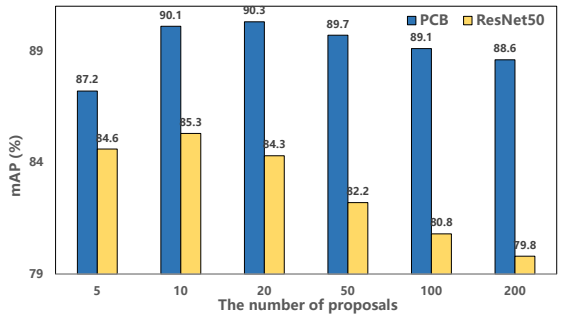

Figure 4. Influence of the number of proposals on CUHK-SYSU with 100 gallery size setting.

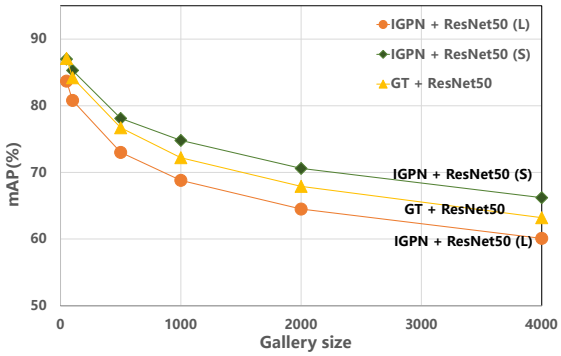

Figure 5. Influence of gallery sizes on CUHK-SYSU dataset. "IGPN + ResNet50 (L)" denotes that we keep 50, 100, 500, 1000, 2000 and 4000 proposals for the gallery sizes of 50, 100, 500, 1000, 2000 and 4000, respectively. "IGPN + ResNet50 (S)" denotes that we keep 10, 10, 25, 25, 25 and 50 proposals for each gallery size. "GT + ResNet50" represents that the Re-id model is tested using ground truth bounding boxes.

ods, our IGPN introduces the appearance information of the query person and outputs the similarity scores of the candidates. Given a set of gallery scene images, our IPGN preserves the person patches similar to the target person by ranking the corresponding similarity scores, which decreases the number of person proposals in detection based methods and decreases the mis-detection rate in search based methods.

Firstly, we conduct a set of experiments on CUHK-SYSU with a gallery size of 100 and list the results in Figure 4. From these results, we can draw the intuitive conclusion that decreasing proposals can improve the performance of person search. This is because a large number of proposals contain more distractors which have a negative influence on the performance of the person Re-id part. However, too few proposals may hurt the final performance because this will lead to a high mis-detection rate.

Then we vary the gallery size from 50 to 4000 to test the influence of gallery sizes. We use the standard ResNet50 as the Re-id model. Intuitively, the task will become more challenging with the gallery size increasing, because more hard scenes and distracting factors will be involved. The results are reported in Figure 5. From the experimental results, we have the following observations: (1) as the gallery size increases, the performance of all the compared methods decreases because more distractors are involved; (2) the proposed IGPN can improve the performance by decreasing the proposals detected from the gallery scene images under all gallery sizes and even outperforms the perfect detector (GT).

4.5. Comparison with State-of-the-Art Methods

In this section, we compare IGPN with several state-of-the-art methods. All the detectors, Re-id models or end-to-end models are built upon ResNet-50, except the detector used in MGTS that is based on VGGNet and denoted as "CNN_v".

Results on CUHK-SYSU. We report the person search results on CUHK-SYSU with 100 gallery size setting in Table 3 and Figure 6, where "CNN" denotes the Faster R-CNN detector. When combined with a separately trained person Re-id model, as shown in Figure 6, our method "IGPN + ResNet50" outperforms the method "Faster-RCNN + ResNet50" (85.3 vs 81.2) and we only need to process 10 proposals in the identity matching process. It is worth noting that QEEPS [16] is a query-guided end-to-end network and achieve very high performance. However, it is hard to use rich information of the human body in an end-to-end network. Therefore, when using the PCB model [20], which can make use of the fine-grained part information of the proposals, our method "IGPN + PCB" outperforms all other competitors. Moreover, although PCB is a stronger baseline than ResNet50, it can still benefit from fewer proposals, which further confirms that our proposed IGPN is compatible with state-of-the-art person Re-id methods.

Table 3. Comparison of performance on CUHK-SYSU with 100 gallery size setting. The number in parenthese denotes the number of proposals we keep for person Re-id.

Method	mAP(%)	top-1(%)
OIM [24]	75.5	78.7
IAN [23]	76.3	80.1
NPSM [14]	77.9	81.2
RCAA [2]	79.3	81.3
CNN_v + MGTS[3]	83.0	83.7
CNN + CLSA [11]	87.2	88.5
Context [27]	84.1	86.5
QEEPS [16]	88.9	89.1
IGPN + ResNet50 (100)	80.8	81.4
IGPN + ResNet50 (10)	85.3	85.7
IGPN + PCB (100)	89.1	90.5
IGPN + PCB (20)	90.3	91.4

Results on PRW. On PRW dataset, we conduct experiments to compare IGPN with the state-of-the-art methods

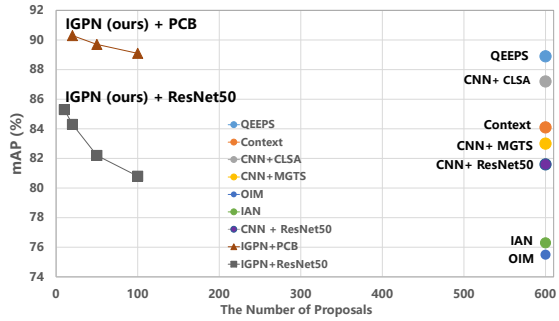

Figure 6. Comparison of performance on CUHK-SYSU with 100 gallery size setting. There is an average of around 6 annotated bounding boxes per image on CUHK-SYSU so for other detection based methods, the number of proposals should be around 600 under 100 gallery size setting.

Table 4. Comparison of performance on PRW. The number in parenthese denotes the number of proposals we keep for person Re-id.

Method	mAP(%)	top-1(%)
OIM [24]	21.3	49.9
IAN [23]	23.0	61.9
NPSM [14]	24.2	53.1
CNN_v + MGTS [3]	32.6	72.1
CNN + CLSA [11]	38.7	65.0
Context [27]	33.4	73.6
QEEPS [16]	37.1	76.7
IGPN + ResNet50 (6k)	41.1	81.2
IGPN + ResNet50 (1.2k)	42.9	82.1
IGPN + PCB (6k)	46.2	86.1
IGPN + PCB (1.2k)	47.2	87.0

and report the results in Table 4. The results show that there is a significant performance gap between PRW and CUHK-SYSU. Our method "IGPN + ResNet50" can outperform all other methods. This is because the gallery set of PRW is very large and contains many people wearing similar clothes. Therefore, a separately trained person Re-id model which can learn more discriminative features of the person can benefit the search task more. When taking IGPN as a detection network, we can still improve the performance by decreasing proposals. There are 25,062 bounding boxes in the testing set of PRW, so for a query person, the person Re-id model needs to process around 25K person patches in other detection based methods. However, thanks to IGPN, the identity matching process is more efficient and we only need to feed 1,200 patches to the Re-id model.

4.6. Runtime Comparison

The inference time of our method and other detection based methods is reported in Table 5. We set the number of proposals fed into the head of our IGPN detector as 32. As our method is instance-guided and the correlation kernels only need to be calculated once for the same query, it is the fastest for one query. Moreover, our method is also more efficient than QEEPS which is also query-guided because QEEPS need to process both query and gallery images at all times.

Table 5. Inference time on CUHK-SYSU with gallery size of 100. * means the work we re-implement in PyTorch.

Method	IGPN(ours)	OIM*	MGTS	QEEPS
1 Query	6s	15s	127s	30s

4.7. Discussion

Compared with other methods, our two-stage method owns the following merits: (a) Different from the conventional detectors in previous two-stage methods, our IGPN can generate high-quality proposals for the identity matching process, which can benefit the overall search performance. (b) Compared with search-based methods, our IGPN can decrease the mis-detection rate by preserve multiple proposals within a scene image if necessary. (c) Our IGPN is more efficient in the proposal generation phase.

There are also some weaknesses. Firstly, IGPN and the Re-id model are trained separately. Secondly, IGPN is instance-guided so it is inefficient for multiple queries if the gallery is shared. In the future work, we will explore to boost our method with multiple queries and train the two parts in an end-to-end fashion.

5. Conclusion

In this paper, we present a novel Instance Guided Proposal Network (IGPN) for person search. Unlike person detection networks in the previous methods, our IGPN can learn the similarity between queries and proposals by leveraging the appearance information of queries, local relations between proposals and global relations from scenes in an end-to-end manner. Thus, we can decrease the proposals fed into the following person Re-id part by keeping the bounding boxes with high similarity scores. We have conducted extensive experiments to evaluate the performance of our model and the experimental results verify its superiority.

Acknowledgements

This work was supported in part by the Major Project for New Generation of AI (No.2018AAA0100400), 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] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Proceedings of the European Conference on Computer Vision*, pages 850–865. Springer, 2016. 2

[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. 1, 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. 1, 2, 3, 5, 7, 8

[4] 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. 6

[5] 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

[6] 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

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 5

[8] 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. 3

[9] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018. 2

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 5

[11] 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. 1, 2, 3, 5, 6, 7, 8

[12] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. *arXiv preprint arXiv:1812.11703*, 2018. 2

[13] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018. 2, 3, 5, 6

[14] 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. 1, 2, 3, 7, 8

[15] Chuanchen Luo, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Spectral feature transformation for person re-identification. *arXiv preprint arXiv:1811.11405*, 2018. 2

[16] 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. 2, 5, 7, 8

[17] 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

[18] 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

[19] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European Conference on Computer Vision*, pages 486–504, 2018. 2

[20] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision*, pages 480–496, 2018. 5, 6, 7

[21] Jack Valmadre, Luca Bertinetto, João Henriques, Andrea Vedaldi, and Philip HS Torr. End-to-end representation learning for correlation filter based tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2805–2813, 2017. 2

[22] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 4

[23] 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, 8

[24] 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, 5, 7, 8

[25] 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

[26] 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

[27] 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, 8

[28] 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

[29] 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. 1, 2, 5

[30] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European Conference on Computer Vision*, pages 101–117, 2018. 2, 3