This CVPR 2020 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# TCTS: A Task-Consistent Two-stage Framework for Person Search

Cheng Wang[1], Bingpeng Ma[1]*, Hong Chang[2,1], Shiguang Shan[2,1,3], Xilin Chen[2,1]
[1]University of Chinese Academy of Sciences, Beijing, 100049, China
[2]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[3]Peng Cheng Laboratory, Shenzhen, 518055, China
wangcheng18@mails.ucas.ac.cn, bpma@ucas.ac.cn, {changhong, sgshan, xlchen}@ict.ac.cn

## Abstract

*The state of the art person search methods separate person search into detection and re-ID stages, but ignore the consistency between these two stages. The general person detector has no special attention on the query target; The re-ID model is trained on hand-drawn bounding boxes which are not available in person search. To address the consistency problem, we introduce a Task-Consist Two-Stage (TCTS) person search framework, includes an identity-guided query (IDGQ) detector and a Detection Results Adapted (DRA) re-ID model. In the detection stage, the IDGQ detector learns an auxiliary identity branch to compute query similarity scores for proposals. With consideration of the query similarity scores and foreground score, IDGQ produces query-like bounding boxes for the re-ID stage. In the re-ID stage, we predict identity labels of detected bounding boxes, and use these examples to construct a more practical mixed train set for the DRA model. Training on the mixed train set improves the robustness of the re-ID stage to inaccurate detection. We evaluate our method on two benchmark datasets, CUHK-SYSU and PRW. Our framework achieves 93.9% of mAP and 95.1% of rank1 accuracy on CUHK-SYSU, outperforming the previous state of the art methods.*

## 1. Introduction

Person search is the extension of person re-identification (re-ID). Person search aims to localize specific targets in the whole scenarios. It can be seen as the organic combination of person detection and re-ID, which meets the practical requirements. It is applicable for many fields, such as video surveillance systems, people finder systems in parks, self-service supermarket and so on. Therefore, more and more researchers put attention on person search task.

*Corresponding Author

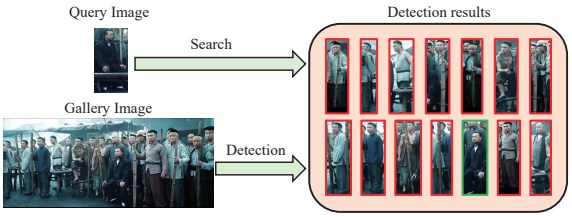


Figure 1. The consistency problem in the detection stage. The red boxes indicates non-query pedestrians and the green box indicates the query target in gallery image. The person detector produces bounding boxes of each pedestrian, which leads to a large gallery size for the re-ID stage.

Deep-learning based person search methods can be divided into two classes. One is *two-step*, which separates the network parameters of the two tasks. In this way, person search is regarded as a sequential procedure with two stages, detection and re-ID. The other is *end-to-end*, which learns a shared feature representation for person detection and person re-ID. The model receives the training signals from the two tasks at the same time. However, the detection task focuses on the commonness of pedestrians while the re-ID task focuses on the uniqueness of pedestrians. There exists a conflict between two tasks in joint learning, which eventually influences the optimization of model. Thus, we adopt the two-step structure in this paper.

Exists two-stage methods fail to notice the consistency requirements between the sub-tasks in person search. For the detection stage, a general person detector is not consistent with the follow re-ID task. The general person detector produces bounding boxes of each pedestrian (Fig. 1), so that the re-ID stage suffers from a large gallery size, which increases the difficulty of recognition. Besides, because the query targets are not received particular attention, the problems of false alarm (on the background) and missed detection on query targets are more critical. This kind of detection error will lead to further matching errors inevitably. Therefore, the detector in person search frame-


Figure 2. The consistency problem in the re-ID stage. The first line are examples from re-ID train set, the second line are corresponding detection results. In actual search process, the detected bounding boxes are more vulnerable to problems of misalignment, occlusions and body part missing.

work should serve for the re-ID stage to recognize the query targets. Training a superior general person detector does not meet the consistency requirement of the re-ID stage. It is worth mentioning that the end-to-end method QEEPS [18] has exploited query information in the detection process. However, the features used in their QRPN are extracted by the base network, which focuses on the commonness of the pedestrians. Their attention mechanism helps suppression background proposals, but be less effective on different pedestrians. Therefore, we need a detector that focuses on the query targets to meet the consistency requirement.

For the re-ID stage, the re-ID model is not consistent with the complex detection results. Compared with the hand-drawn bounding boxes, the detected bounding boxes are more vulnerable to problems of misalignment, occlusions and body part missing (Fig. 2). Since the person search datasets provide identity annotations only on hand-drawn boxes, exists two-step methods train the re-ID model on cropped ideal person images. However, the inputs of the re-ID stage are detected bounding boxes in actual practice. Therefore, there is an inconsistency between the re-ID stage with the detection outputs. As a result, the re-ID model outputs error recognition result in inaccurate detected bounding boxes. This consistency problem in re-ID stage degrades the search performance and limits the practicability. Han *et al*. [8] also notice the performance degradation of the re-ID model on detection boxes. They refine the detection boxes by a learnable affine transformation. For one thing, this method is of no help to missed detection. For another thing, the upper bounding of the refinement is approaching the ground truth, so the performance improvement is constrained. Therefore, the consistency problem is an unsolved issue in person search.

In this paper, we propose a novel two-stage framework to eliminate the inconsistencies in the detection and re-ID stage respectively. For the detection stage, we propose an identity-guided query (IDGQ) detector to produce query-like bounding boxes. Considering the commonness information in general person detector does not differentiate query targets from other pedestrians, the IDGQ detector has an auxiliary identity branch to compute query similarity scores for proposals in the whole scenarios. By considering both query-like similarity score and foreground score, IDGQ outputs more accurate query-like bounding boxes and fewer non-query bounding boxes. In order to improve the discrimination of the identity branch, we propose a novel classification loss. For each gallery image, by minimizing the probability of the labeled examples being recognized as different pedestrians in the same image, the auxiliary branch can output superior query-like similarity score for query proposal than other proposals. Thus, it further focuses IDGQ detector on query-like proposals.

For the re-ID stage, we propose a Detection Results Adapted (DRA) re-ID model. In order to make the re-ID model adapted to the detection results, we provide a mixed train set to the DRA model, which contains hand-drawn and detected bounding boxes at the same time. Specifically, the identity labels of detected boxes are predicted based on the ground truth boxes according to the IoU overlap. These annotated detected bounding boxes will be used to train the DRA model. Further, considering that the quality of the detected examples is uneven, and the majority of the mixed set is the easy accurate-detected example, we propose an example reweighting algorithm. This algorithm automatically reweights examples with consideration of quality and hardness. Thus, the DRA has both steady convergence rate and robustness to inaccurate detection.

The experimental results on two benchmark datasets show that our detector achieves a higher recall and accuracy on query targets. The re-ID part achieves a higher matching performance when testing on the detection results. Our framework achieves 93.9% of mAP and 95.1% of rank1 accuracy on CUHK-SYSU, outperforming the state of the art of two-stage methods.

## 2. Related Work

**Person Search**. The person search task aims to locate and identify a person at the same time. A simple solution is to combine person detection with person re-ID sequentially, but the search performance will be restricted to the two components, and the setting of re-ID is different from person search. In 2014, Xu *et al*. [24] introduce the concept of person search and propose a model based on hand-crafted features using the strategy of sliding windows.

Recently, with the rise of deep learning methods and the large-scale person search datasets, several frameworks are proposed. Some works attempt to solve location and re-ID in an end-to-end way. Xiao et al. [23] propose an end-to-end person search model, and jointly train person detection network and re-ID network. Liu et al. Yan et al. [25] propose a region-based feature learning model, and build a graph for pedestrians in the different input images to learn

contextual information.

Other works solve person search by two steps, i.e. training two parameter independent models for detection and re-ID. Zheng et al. [32] test various combinations of detectors and recognizer, and propose a CWS way to transfer classification confidence from the detector to the re-ID network. These two-step methods [32, 3] improve the search performance by feeding more information from the detector or input images to the recognizer.

**Person search and re-ID datasets**. Though the study of person re-ID [27, 1] made a great progress in recent years, most of recent image-based re-ID datasets (such as VIPeR [7], PRID2011 [10], CUHK01 [14], CUHK02 [13], iLIDS-VID [33] and Duke-MTMC [20]) draw gallery bounding boxes by using human labor. These ideal bounding boxes are unavailable in practical applications. Therefore, some datasets use bounding boxes produced by person detector, such as CUHK03 [15], Market1501 [30] and MARS [29]. It is worth mentioning that CUHK03 also provide a hand-drawn version. Many experiment results show that the using detected bounding boxes leads to an inferior re-ID accuracy than the hand-drawn version. Therefore, false detection and misalignment are actually a critical problem in practice.

As for person search datasets, both PRW and CUHK-SYSU provide only ground truth bounding boxes for the training of the re-ID part. In the testing phase, the pedestrian bounding boxes are all produced by the detector. To the best of our knowledge, all of the previous person search works which have explicit re-ID process train the re-ID model the ground truth bounding boxes. Therefore, there is a gap between the training phase and the testing phase of the re-ID part.

# 3. A Task-Consistent Two-stage Framework for Person Search

In this section, we present our task-consistent person search framework TCTS. As shown in Fig. 3, given a query image and a gallery image, the IDGQ detector produces proposals and computes the query similarity score and foreground score for each proposal. After selection, the outputs of the IDGQ detector are query-like boxes, which are then fed into the DRA net.

## 3.1. Identity-Guided Query Detector

IDGQ detector produces query-like bounding boxes for the re-ID stage by an auxiliary identity branch. As illustrated in Fig. 3, the IDGQ detector has a shared base network and two branch networks. The first branch is the auxiliary identity branch. In order to introduce the identity information into the detector, we train the auxiliary branch by a classification loss, called IDGQ loss. In this way, the auxiliary branch can compute identity similarity scores between query target and proposals. In the other branch, we keep the standard head of faster R-CNN [19], because the binary classifier can output accurate foreground scores. These two scores guide the IDGQ detector to keep query-like and foreground proposals.

Specifically, given a cropped query image and a gallery image, the base network and RPN first produce proposals on the gallery image. Then, the identity branch extracts features for query and proposals, donated by $\mathbf{x_q}$ and $\mathbf{x_{g_i}}$ respectively. The identity similarity scores are computed by a similarity function $S(\cdot)$. At the same time, the detection branch outputs the foreground score of proposal $s_{f_i}$. The final score for proposal $i$ is calculated by:

$$s_i = S(\mathbf{x_q}, \mathbf{x_{g_i}}) \times s_{f_i} \tag{1}$$

After then, only proposals with high query similarity and foreground scores will be kept. The query-like bounding boxes are consistent with the re-ID task.

It is worth discussing how to design a suitable loss function for it. The target of the identity branch is to output superior query-like similarity scores for query proposals than other proposals. Therefore, it has some differences with the traditional classification or re-ID task. Next, we first introduce two common classification losses. Then, we derive our proposed IDGQ loss.

Softmax loss is widely used in the classification task. For an example $x_i$, Softmax loss uses a Softmax function to calculate the probability on each class and use a cross-entropy loss function to optimize the log-likelihood of each class in probability space. OIM loss is proposed for person search task in [23]. Unlike the Softmax loss, the OIM loss stores a feature center for each person. The center is modified in every iteration by a weighted sum operation. Based on this parameter-free structure, the unlabeled identities are exploited as the form of a circular queue. The probability of $x_i$ being recognized as each identity is calculated by a Softmax function.

We have not directly adopted these classification losses for two reasons. For one thing, similar examples in the same image should be discriminated, while the false positive on the similar examples in different images are acceptable. Therefore, the solution obtained by traditional classification losses may be not optimal. For another thing, the unlabeled examples are not fully exploited in these two losses. Though OIM loss also takes into consideration unlabeled identities, the length of the circular queue is an artificial parameter. If the length is too large, the primitive features in the circular queue are outdated to represent unlabeled identities. If small, the optimizing direction and solution are changed significantly in different mini-batch.

We propose an IDGQ loss to improve the accuracy of the query similarity score in the IDGQ detector. For each

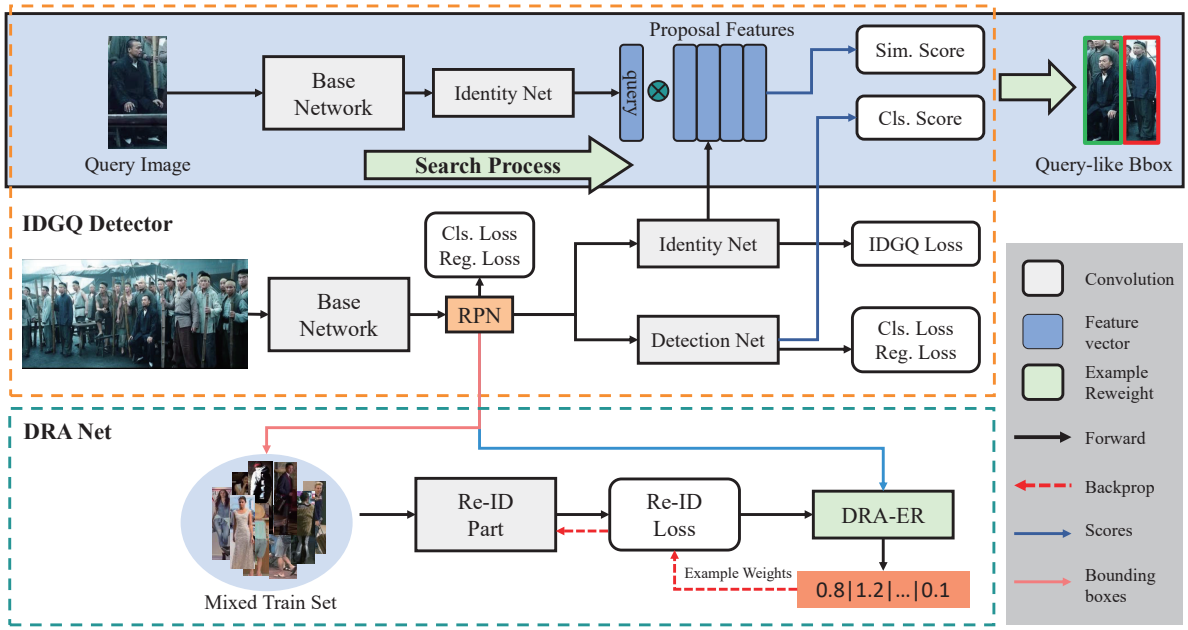


Figure 3. The illustration of our proposed task-consistent two-stage framework. The orange box is the IDGQ detector. In IDGQ, the shared base network extracts common features for detection and identity branches. Then region proposal network(RPN) generates proposals on the feature maps. The identity branch is trained by our proposed IDGQ loss. In the search process, the detection branch and identity branch output foreground scores and query similarity scores, respectively. The green box is our proposed DRA re-ID model. DRA is trained on the mixed train set containing both hand-drawn and detected boxes. In the backward, DRA-ER reweights examples by weight factors.

labeled example, the IDGQ loss pulls the positive examples from the different images closer, so that the images from the same people can receive a high similarity score. Besides, the IDGQ loss pushes the example away from the negative examples (including unlabeled examples) in the same images, which reduce the similarity between different people. Further, in order to effectively exploit unlabeled identities, our IDGQ loss learns a variable number of centers for unlabeled samples.

We store the feature centers $c_i$ of each identity in the memory. For a labeled example $x_i$ in a training image, we construct a reference list $\mathbf{R}$, which contains the feature center $c_{y_i}$, the other labeled examples $\{x_j|j \neq i\}$ in the same image and all of the unlabeled example centers $\mathbf{u}$. the probability of example $x_i$ being recognized as the identity $i$ is:

$$p_i = \frac{exp(c_i^T x_i/\tau)}{\sum_{v \in \mathbf{R}} exp(v^T x_i/\tau)} \tag{2}$$

where $\tau$ is the temperature coefficient. The probability of example $i$ being recognized as the unlabeled center $k$ is:

$$p_k = \frac{exp(u_k^T x_i/\tau)}{\sum_{v \in \mathbf{R}} exp(v^T x_i/\tau)} \tag{3}$$

IDGQ loss optimizes the log-likelihood of its center. The loss function $L_{IDGQ}$ is:

$$L_{IDGQ} = -log(p_i) \tag{4}$$

The unlabeled centers are updated when the inputs are unlabeled examples. We maintain a center list for unlabeled examples, which contains a representative feature and a member list. In the begining, the number of centers is the number of unlabeled examples. Given an unlabeled example $x$, the maximum of $p_k$ is represented as $p_{k^*}$. The unlabeled centers are updated as follow:

$$\begin{cases} u_{k^*} = \alpha u_{k^*} + (1-\alpha)x, & \text{if } p_{k^*} > p_i; \\ u_{U+1} = x, & \text{if } p_{k^*} < p_i. \end{cases} \tag{5}$$

### 3.2. Detection Results Adapted Re-ID Model

In order to train a DRA re-ID model, we construct a mixed train set containing hard-drawn and detected bounding boxes. The detected bounding boxes on the whole scenes are annotated by a pre-trained detector. We predict the identity label of each detected bounding box by the IoU overlap with ground truth. It is worth mentioning that we remove all unlabeled pedestrians from the train set, because the IGDQ detector can effectively suppress the unlabeled pedestrian in advance. In this way, the re-ID train set has 15,085 hand-drawn bounding boxes and 15,198 detected ones. Because there may be several detected boxes on the

same person, the number of the detected boxes is larger than hand-drawn ones. Besides, considering the missed detection problem in the pre-trained detector, there are some identities never appear in the detected part of the train set.

After investigation on the mixed train set, we divide training examples into three types: *accurate* bounding boxes, *misaligned* bounding boxes and *distractors*. *Accurate* includes all of the hand-drawn bounding boxes and some detected bounding boxes which have a large IoU overlap with corresponding ground truth boxes. These examples ensure the convergence rate of the model, but too many easy examples can lead the model to fit these ideal bounding boxes. *Misaligned* and *distractors* have misalignment or human part missing. The difference is that *misaligned* bounding boxes reserve the main focus on corresponding pedestrians. Therefore, these examples are of benefit to train a consistent re-ID model to the detection results. By contrast, *distractors* miss critical details of pedestrians. These *distractors* examples bring wrong gradients direction and lead to a suboptimal model. In a word, the DRA need to focus on different train data at different time.

Example reweight can meet the changing need of DRA by adjusting the importance of examples. We propose an Examples Reweight algorithm for Detection Results Adapted Re-ID, called DRA-ER. We introduce a weighted factor $w$ to determine the example weights. In cross-entropy loss, it is represented as:

$$L_w(x_i) = -w_i log p_{y_i} \tag{6}$$

To solve the excess easy examples, we design a hardness factor $w_h$ to balance the importance of hard and easy examples. The well-trained easy examples are down-weighted so that they draw less attention. Because the distractors are usually hard to be recognized, they have a large $w_h$, too.

Formally, given a labeled example $x_i$, the probability of being recognized as the true class is $p_{y_i}$. The hardness factor $w_h$ is defined as:

$$w_{h_i} = exp((1 - p_{y_i})/T_1) \tag{7}$$

where $T_1$ is the temperature coefficient. When $p_{y_i} \to 1$, the hardness factor achieves a minimum. At the begin, all examples have a small $p_{y_i}$, so the hardness factor $w_q \to 1$ for each example. After the model converges on hand-drawn bounding boxes, the hand-drawn and well-detected examples are down-weighted.

In order to further down-weight the distractor proposals, a quality factor $w_q$ is added to measure the quality of proposals. For detected bounding boxes, we can compute the Intersection over Union (IoU) with the ground truth. In this way, the hand-drawn bounding boxes have $IoU = 1$. The quality factor of the proposal is defined as:

$$w_{q_i} = 1 - \frac{2}{1 + exp((I_i/I^* - 1)/T_2)} \tag{8}$$

where $T_2$ is the temperature coefficient, $I_i$ indicates the IoU between this example and its ground truth, $I^*$ is the threshold to determinate the positive pedestrian example.

The final weight is normed to keep a steady learning rate:

$$w_i = \frac{N w_{h_i} w_{q_i}}{\sum_j^n w_{h_j} w_{q_j}} \tag{9}$$

# 4. Experiments

In this section, we conduct experiments on the two benchmark datasets, CUHK-SYSU and PRW. We compare TCTS with the state of the art methods. Besides, we report the ablation study results of each component in TCTS.

## 4.1. Benchmark

**CUHK-SYSU** [23]. The data come from two sources, street snap, and movie screenshot. For street snap, images are captured by using a hand-held camera in a town. The movies and TV dramas screenshots provide more diversified scenes and different camera states. They cover a wide range of perspectives, lighting, resolutions, occlusions, and background conditions. The dataset contains 18,184 pictures and 8,432 identities. 96,143 pedestrian boundaries are marked in total, and different identities are assigned to people matched in different scenes. The train set has 11,206 images with 5,532 identities, and the test set has 6,978 images with 2,900 identities.

**PRW** [32]. The video frames are captured by 6 cameras in Tsinghua university. All pedestrian bounding boxes are marked manually. A total of 11,816 frames are collected and 4,310 pedestrian bounding frames are obtained. If the pedestrian appears in the MarKet-1501 dataset, a positive identity label is added. Thereby, 34,304 pedestrian bounding frames are labeled with identity labels (from 1 to 932), and the remaining pedestrian bounding frames are labeled with "-2" labels. These people with "-2" labels are not used in the testing of person re-ID, but can be potentially used in the training.

## 4.2. Network Details and Model Training

**Network Details**. Our proposed IDGQ is implemented based on Faster R-CNN[6]. We use ResNet-50 as our backbone network. The base network has four blocks from conv1 to conv4. The detection and identity branches are built upon conv4_6. The identity branch adopts conv5_1 in ResNet-50, and the detection branch adopts conv5. The features of the identity branch are projected to a L2-normalized 256-dimensional subspace, and features are projected to a

Table 1. Comparison of mAP(%) and rank-1 accuracy(%) with the state-of-the-art on CUHK-SYSU. The gallery size is set to 100.

| | Methods | mAP | rank-1 |
|---|---|---|---|
| end-to-end | OIM[23] | 75.5 | 78.7 |
| | IAN[21] | 76.3 | 80.1 |
| | NPSM[17] | 77.9 | 81.2 |
| | RCCA[2] | 79.3 | 81.3 |
| | QEEPS[18] | 84.4 | 84.4 |
| | GRAPH[25] | 84.1 | 86.5 |
| two-step | ACF[26]+DSIFT[28]+Euclidean | 21.7 | 25.9 |
| | ACF+DSIFT+KISSME[11] | 32.3 | 38.1 |
| | ACF+LOMO[16]+XQDA[16] | 55.5 | 63.1 |
| | ACF+IDNet[23] | 56.5 | 63.0 |
| | CCF[26]+DSIFT+Euclidean | 11.3 | 11.7 |
| | CCF+DSIFT+KISSME | 13.4 | 13.9 |
| | CCF+LOMO+XQDA | 41.2 | 46.4 |
| | CCF+IDNet | 50.9 | 57.1 |
| | CNN+DSIFT+Euclidean | 34.5 | 39.4 |
| | CNN+DSIFT+KISSME | 47.8 | 53.6 |
| | CNN+LOMO+XQDA | 68.9 | 74.1 |
| | CNN+IDNet | 68.6 | 74.8 |
| | CNN+MGTS[3] | 83.0 | 83.7 |
| | CNN+CLSA[12] | 87.2 | 88.5 |
| | Re-ID Driven[8] | 93.0 | 94.2 |
| | TCTS | **93.9** | **95.1** |

L2-normalized 1,024-dimensional subspace in the detection branch. Obviously, the identity branch is a light network with less computational efforts. The positive score threshold is set to $0.4 * max(s_q)$ in IDGQ, where $s_q$ is the query similarity score. For the DRA re-ID model, we construct a re-ID baseline with batch normalize layers based on ResNet-50.

**Model Training**. For the detection stage, we first train the network without the identity branch on benchmark datasets, then initialize IDGQ with this pre-trained model. Especially, the first stage of the base network is fixed after initialized. We train the pre-trained detection model for 50K iterations using Stochastic Gradient Descent (SGD) algorithm with momentum set to 0.9. After initialized with the pre-trained model, we train the whole network for 50K iterations using the SGD algorithm. For both pre-training and training, the learning rate is $0.001$ for the first 30K iterations, and decays to $0.0001$. The batch size is set to 2 because the detector part has a lot of intermediate results. To avoid potential problems of local optima and slow convergence, we use the average of the losses from the last 100 iterations. The temperature coefficient is set to $1/50$. The center update coefficient $\alpha$ is set to 0.5.

For the re-ID stage, the DRA model is pre-trained on ImageNet. All the training images are resized to $256 \times 128$. The batch size is set to 128. We adopt the Adam algorithm with default hyper-parameters set in PyTorch. The initial learning rate is $3.5e-4$. Besides, we adopt label smooth, random erasing [34] and triplet loss [9] to improve the search performance. The temperature coefficients $T_1$ and $T_2$ are set to $0.3$ and $5$ respectively.

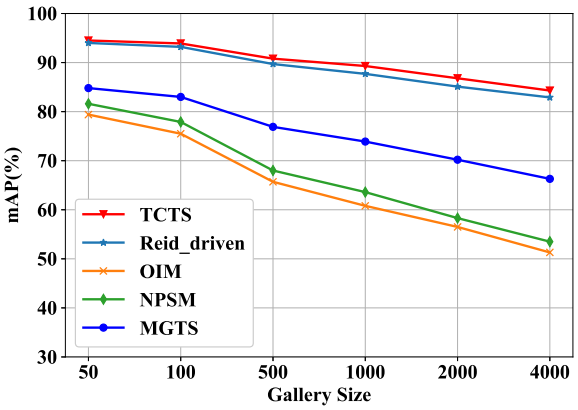


Figure 4. The mAP performance under different gallery size on CUHK-SYSU.

### 4.3. Comparison with State-of-the-art Methods

We compare our TCTS framework with several state-of-the-art methods. Some of these are methods solve person search task by two steps. [32] uses off-the-shelf detectors, including ACF[4], CCF[26] and CNN detector, and hand-crafted descriptors, including DSIFT[28] and LOMO[16], and distance metrics, including Euclidean, KISSME[11] and XQDA[16]. [3], [12] and [8] use Faster-RCNN[19] as detector and proposed re-ID model to solve person search. Other methods are joint learning methods, including OIM[23], IAN[21], NPSM[17] and RCCA[2].

In order to evaluate the model robustness to the variation of gallery size, we repeat the experiment on CUHK-SYSU dataset under the different gallery sizes ([50, 100, 500, 1000, 2000, 4000]), and make a comparison to other deep-learning based methods. In Fig. 4, our method outperforms other methods under 6 kinds of gallery sizes, which proves the strong robustness of TCTS.

**Results on CUHK-SYSU**. Table 1 reports the performance of TCTS, and gives comparisons with the state of the art approaches for the CUHK-SYSU dataset. In the table, the "CNN" detector is the Faster R-CNN[6] with a ResNet-50 backbone network. "GT" indicates directly using the ground truth bounding boxes as the detector results. "IDNet" learns discriminative re-ID feature representations by using Softmax loss to train a classifier for the person with different identities. "MGTS" indicates a Mask-guided Two-Stream CNN Model. "CLSA" indicates a Cross-Level Semantic Alignment deep learning approach. The training scheme can be seen in [22]. The gallery size is set to 100.

From Table 1, we can draw the following conclusions:

Table 2. Comparison of mAP(%) and rank-1 accuracy(%) with the state-of-the-art on PRW. The gallery size is set to 6,112.

| | Methods | mAP | rank-1 |
|---|---|---|---|
| end-to-end | OIM[23] | 21.3 | 49.9 |
| | IAN[21] | 23.0 | 61.9 |
| | NPSM[17] | 24.2 | 53.1 |
| | GRAPH[25] | 33.4 | 73.6 |
| | QEEPS[18] | 37.1 | 76.7 |
| two-step | ACF[26] + LOMO[16] + XQDA | 10.3 | 30.6 |
| | ACF + IDE$_{det}$[32] | 17.5 | 43.6 |
| | ACF + IDE$_{det}$ + CWS[32] | 17.8 | 45.2 |
| | DPM[5] + LOMO + XQDA | 13.0 | 34.1 |
| | DPM + IDE$_{det}$ | 20.3 | 47.4 |
| | DPM + IDE$_{det}$ + CWS | 20.5 | 48.3 |
| | LDCF + LOMO + XQDA | 11.0 | 31.1 |
| | LDCF + IDE$_{det}$ | 18.3 | 44.6 |
| | LDCF + IDE$_{det}$ + CWS | 18.3 | 45.5 |
| | CNN+MGTS[3] | 32.6 | 72.1 |
| | CNN+CLSA[12] | 38.7 | 65.0 |
| | Re-ID Driven[8] | 42.9 | 70.2 |
| | TCTS | **46.8** | **87.5** |

(1) Among deep-learning based methods, the two-step frameworks outperform the end-to-end ones, and the advantages are becoming more and more obvious. The main reason is that the conflict between two tasks in joint learning influences the optimization of model. For example, our TCTS outperforms the best end-to-end method "QEEPS" by 9.8% on mAP and 8.6% on rank-1 accuracy. It is worth mentioning that "QEEPS" proposes a QRPN to produce query-like proposals like us. In QRPN, the query guidance is achieved by channel attention base on query features. Because the query features are extracted by the base network but not its ClsIdenNet, we argue that IDGQ can pay more attention to query target than QRPN.

(2) Both the mAP scores and rank-1 accuracy of TCTS are higher than comparative methods. In particular, it outperforms the best two-step method "Re-ID Driven" [8] by 0.9% on both the mAP performance and rank-1 accuracy. In [8], they argue that the detected bounding boxes may be suboptimal for the following re-ID task, so they propose a learnable refinement network for providing refined detection boxes. Compared with "Re-ID Driven", our IDGQ can also reduce missed detection on query target, and the DRA model is more robust to unavoidable detection errors. These advantages bring more performance improvement.

**Results on PRW**. We also compare the different methods on the PRW dataset. The evaluation results are shown in Tab. 2. In the table, "IDE$_{det}$" [31] indicates first training an R-CNN model on PRW, then fine-tuning the R-CNN model with the IDE method. "CWS" indicates incorporating detection confidence into the similarity measurement. In this experiment, the gallery size of the test dataset is set to 6,112.

Table 3. Evaluating effectiveness of IDGQ detector and our proposed IDGQ loss on CUHK-SYSU. The number of ground truth boxes is 8,340.

| Methods | Boxes Num | Recall | mAP | rank-1 |
|---|---|---|---|---|
| Faster R-CNN | 54,658 | 96.9 | 91.4 | 92.4 |
| IDGQ | 27,563 | **98.2** | **93.9** | **95.1** |
| IDGQ(OIM) | **27,332** | 97.3 | 92.0 | 92.9 |

Compared with other methods, our TCTS improves about 4% on the mAP and 17.3% on the rank-1 accuracy. In the test set of PRW, the full set is used as gallery, so there is a tremendous number of detected bounding boxes. In the IDGQ detector, the identity branch suppresses non-query proposals so that the gallery size of the re-ID stage is under control. Besides, each query target appears around 50 times in gallery images, which brings notable inter-class variations. The result of mAP shows the identity branch can output accurate query similarity scores in this situation, which proves the effectiveness of the IDGQ loss.

### 4.4. Ablation Study

To validate the effectiveness of each component of TCTS, we implement several ablation experiments on the CUHK-SYSU dataset. The gallery size is set to 100 for all experiments.

**IDGQ and IDGQ loss.** In TCTS, we use the IDGQ detector to produce more accurate query-like bounding boxes and less non-query bounding boxes. In this experiment, we evaluate the effectiveness of the IDGQ detector and IDGQ loss. In Tab. 3. "recall" indicates only the recall on query targets. "Boxes Num" is the number of detected boxes for all query targets. The positive score threshold is set to 0.5 in Faster R-CNN, and is $0.4 * max(s_q)$ in IDGQ, where $s_q$ is the query similarity score.

From the table, we observe that IDGQ has a higher query recall than faster R-CNN. The reason is that some query proposals with low foreground scores receive high query similarity scores, so their final scores will be above the threshold. Besides, the improvement in query recall is achieved by even less bounding boxes. In another word, besides the query target, faster R-CNN produces around $5.5$ non-query boxes. The number is reduced to $2.3$ in IDGQ. It demonstrates that IDGQ effectively focuses on query-like proposals. Besides, the 39,514 unlabeled images are clustered into 3,278 centers. It proves the effectiveness of the learning of variable unlabeled centers.

If we replace IDGQ loss with OIM loss, the search performance decreases from 93.9% to 92.0% on mAP. It verifies that IDGQ loss obtains a more optimal solution than OIM loss.

**Comparison between Hardness Factor and Focal Loss.** As shown in Fig. 5, our hardness weight ($T = 0.3$)

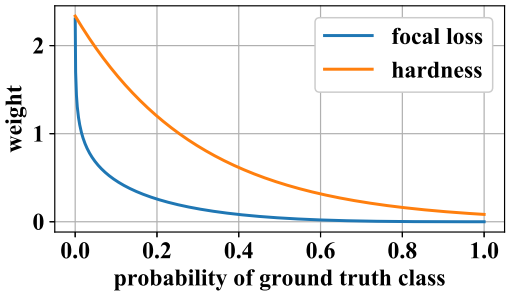


Figure 5. The weight varied with the probability of ground truth class for our hardness factor and the focal loss.

Table 4. Evaluating effectiveness of DRA-ER algorithm.

| Train set | mAP | rank-1 |
|---|---|---|
| Hand-drawn | 93.1 | 94.0 |
| Detected | 82.9 | 83.8 |
| Mixed | 86.9 | 88.2 |
| Mixed+ER | **93.9** | **95.1** |

has a similar change curve with focal loss ($\gamma = 2, \alpha = 0.25$). The curve of the hardness weight uses relative values due to the normalization in Eq.9. Compared with focal loss, our proposed hardness factor is more suitable for the person search training sets. The hardness factor reduces losses of easy examples to a small value, but no to 0 as the focal loss does. Considering the easy positive examples have different identities and the scale of the training set is limited, the hardness factor keeps the variety of training examples.

In our experiments, if we replace the hardness factor with the weight factor in the focal loss, the model only achieves similar performance as baseline trained on the hand-drawn training set.

**Effectiveness of the DRA-ER algorithm.** We propose a DRA-ER algorithm to automatically reweight examples by the need of the model. In this experiment, we first verify the uneven quality problem in detected bounding boxes, and then evaluate the effectiveness of the example reweight algorithm in the mixed set. "Hand-drawn" indicates using hand-drawn boxes only, "Detected" indicates using detected boxes with predicted identity labels only, "Mixed" indicates using our constructed mixed train set.

The results are reported in Tab. 4. We observe that "Detected" achieves a lower search performance than "Hand-drawn". Though the detected boxes make the training and testing of the re-ID model consistent, the uneven quality problem has more significant influences on the performance. We also observe that "Hand-drawn" also outperforms "Mixed". It shows that using the mixed set can not directly bring performance improvement, and it even influences the convergence of the model. In the last row of the table, we show that adopting the DRA-ER algorithm on the mixed set can achieve better performance than "Hand-drawn". In the early stage of training, hand-drawn and accurate detected boxes are up-weighted due to the high quality factor. In the later stage of training, DRA-ER balances the importance between easy examples and some hard detected boxes. Besides, the low quality detected boxes have almost no effect on the training.

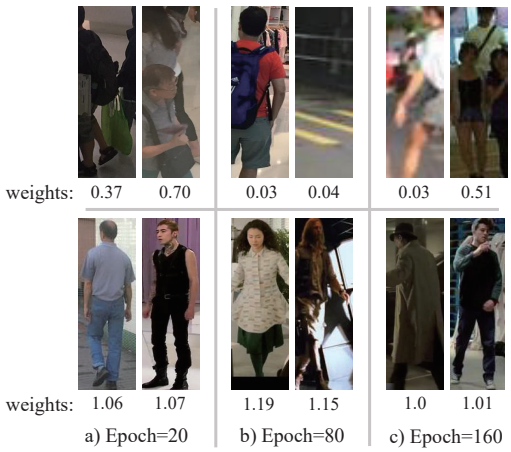


Figure 6. The illustration of up-weighted and down-weighted examples in different training stages.

In Fig.6, we illustrate the weights changing by visualizing some typical examples and their weights. The first row is the down-weighted examples, and the section row is the up-weighted examples. We observe that the low-quality detected examples with part missing or false alarm problems are down-weighted throughout the training. The attention of DRA-ER changes from accurate boxes to the hard detected boxes. In the epoch 160, these two kinds of boxes finally have similar weights. It demonstrates that the DRA model is adapted to the detected boxes.

## 5. Conclusion

In this work, we point out the consistency problem in exists two-step person search framework. To address that, we propose a TCTS framework that has an IDGQ detector and a DRA re-ID stage. The IDGQ detector can effectively produce query-like bounding boxes, which achieves a higher query recall and reduces the number of bounding boxes. The DRA achieves a better performance on detected results, which is attributed to the detected train data and the DRA-ER algorithm. TCTS achieves the state of the art performance on two person search benchmark datasets, CUHK-SYSU and PRW.

**Acknowledgement.** This work is partially supported by Natural Science Foundation of China (NSFC): 61732004, 61876171 and 61976203.

## References


[1] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *IEEE Computer Vision and Pattern Recognition*, 2017.

[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 *European Conference on Computer Vision*, 2018.

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

[4] 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), 2014.

[5] 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), 2010.

[6] Ross Girshick. Fast r-cnn. In *IEEE International Conference on Computer Vision*, 2015.

[7] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, 2008.

[8] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-id driven localization refinement for person search. In *IEEE International Conference on Computer Vision*, 2019.

[9] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[10] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian Conference on Image Analysis*, 2011.

[11] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *IEEE Computer Vision and Pattern Recognition*, 2012.

[12] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In *European Conference on Computer Vision*, 2018.

[13] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *IEEE Computer Vision and Pattern Recognition*, 2013.

[14] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *Asian Conference on Computer Vision*, 2012.

[15] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Computer Vision and Pattern Recognition*, 2014.

[16] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Computer Vision and Pattern Recognition*, 2015.

[17] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. In *IEEE International Conference on Computer Vision*, 2017.

[18] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. *arXiv preprint arXiv:1905.01203*, 2019.

[19] 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*, 2015.

[20] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[21] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. Ian: the individual aggregation network for person search. *arXiv preprint arXiv:1705.05552*, 2017.

[22] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *IEEE Computer Vision and Pattern Recognition*, 2016.

[23] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *IEEE Computer Vision and Pattern Recognition*, 2017.

[24] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. Person search in a scene by jointly modeling people commonness and person uniqueness. In *ACM international conference on Multimedia*, 2014.

[25] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *IEEE Computer Vision and Pattern Recognition*, pages 2158–2167, 2019.

[26] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z. Li. Convolutional channel features. In *IEEE International Conference on Computer Vision*, 2015.

[27] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification. In *European Conference on Computer Vision*, 2018.

[28] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *IEEE Computer Vision and Pattern Recognition*, 2013.

[29] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, 2016.

[30] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, 2015.

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

[32] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *IEEE Computer Vision and Pattern Recognition*, 2017.

[33] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *British Machine Vision Conference*, 2009.

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