Sequential End-to-end Network for Efficient Person Search

Zhengjia Li[1,2], **Duoqian Miao**[1,2*]
[1]Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
[2]Key Laboratory of Embedded System and Service Computing, Ministry of Education, Shanghai 201804, China
zjli1997@tongji.edu.cn, dqmiao@tongji.edu.cn

Abstract

Person search aims at jointly solving Person Detection and Person Re-identification (re-ID). Existing works have designed end-to-end networks based on Faster R-CNN. However, due to the parallel structure of Faster R-CNN, the extracted features come from the low-quality proposals generated by the Region Proposal Network, rather than the detected high-quality bounding boxes. Person search is a fine-grained task and such inferior features will significantly reduce re-ID performance. To address this issue, we propose a Sequential End-to-end Network (SeqNet) to extract superior features. In SeqNet, detection and re-ID are considered as a progressive process and tackled with two sub-networks sequentially. In addition, we design a robust Context Bipartite Graph Matching (CBGM) algorithm to effectively employ context information as an important complementary cue for person matching. Extensive experiments on two widely used person search benchmarks, CUHK-SYSU and PRW, have shown that our method achieves state-of-the-art results. Also, our model runs at 11.5 fps on a single GPU and can be integrated into the existing end-to-end framework easily.

Introduction

Pedestrian detection (Girshick et al. 2014; Girshick 2015; Ren et al. 2015) aims at detecting the bounding boxes (BBoxes) of all people in the image. Person re-identification (re-ID) (Yang et al. 2017; Zhao et al. 2017; Wang et al. 2019; Fu et al. 2019; Hao et al. 2019; Zhao et al. 2020) is used to match the interested person with hand-cropped person images. Although these two fields are widely studied in recent years, they can not be directly applied to real-world applications due to their limited functionality. To close the gap, Xu et al. introduce person search task which aims at locating a target person in the scene image (Xu et al. 2014). Person search can be seen as a combination of pedestrian detection and person re-ID. It has broad application prospects in video surveillance, finding lost children, and self-service supermarket, *etc*.

As illustrated in Figure 1, existing works divide the task into generating BBoxes of all people in the image and person re-ID. They either tackle the problem separately with two independent models (*two-stage* methods) or jointly with a multi-task model (*end-to-end* methods).

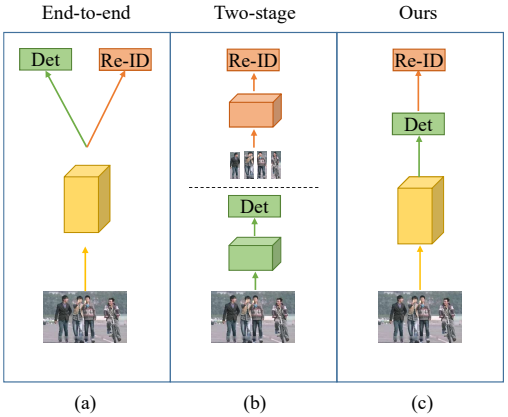

Figure 1: Comparison of three methods for person search. (a). Existing end-to-end framework. (b). Existing two-stage framework. (c). Ours.

For end-to-end methods (Xiao et al. 2017, 2019; Munjal et al. 2019), they design a multi-task framework based on Faster R-CNN (Ren et al. 2015). A Region Proposal Network (RPN) is built to generate region proposals, which are then fed into the subsequent parallel detection and re-ID branches. However, these features extracted by the network come from low-quality proposals rather than detected accurate BBoxes. Although these inferior features have little impact on the coarse-grained classification task, they will significantly reduce the performance of the fine-grained re-ID task. This problem is caused by the parallel structure of Faster R-CNN. Because detection and re-ID are processed at the same time, the accurate BBoxes are not available before extracting re-ID features. For two-stage methods, there is no such problem, because detection and re-ID are tackled sequentially with two separate models. However, they are time-consuming and resource-consuming.

Motivated by the above observations, we propose a Sequential End-to-end Network (SeqNet) illustrated in Figure 1 (c) to extract high-quality features. Specifically, detection and re-ID share the stem representations, but solved with two head networks sequentially. Compared with baseline,

*Corresponding author.

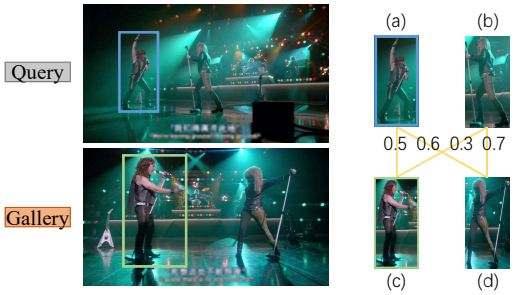

Figure 2: An example of matching query image and gallery image. Blue box denotes the query person and green box denotes the ground truth. The number attached to each yellow line represents the similarity between the two people connected by the line.

our model employs an extra Faster R-CNN head as an enhanced RPN to provide high-quality BBoxes. Then an unmodified baseline head is used to extract the discriminative features of these BBoxes. At test time, the non-maximum suppression (NMS) is applied to remove redundant BBoxes before re-ID stage for efficiency. Moreover, to improve the classification ability of baseline head for high Intersection over Union (IoU) samples, we adopt the more reliable classification result of detection head. In general, our SeqNet not only inherits the sequential process of two-stage methods, which can provide accurate BBoxes for re-ID stage, but also retains the end-to-end training fashion and efficiency.

Another challenge for person search is how to utilize context information to perform more robust matching. As shown in Figure 2, given a query person (a), (c) is the corresponding ground truth, but the (d) with largest similarity (0.6) will be mistakenly predicted as top-1 result. If context information (b) is taken into consideration, to maximize the total similarity, the optimal matching should be $(a) \leftrightarrow (c), (b) \leftrightarrow (d)$. In this way, the wrong prediction (d) can be revised to (c). Inspired by this, we design a Context Bipartite Graph Matching (CBGM) algorithm to exploit context information as a complement to individual feature. Specifically, we treat all people in the query image and each gallery image as two sets of vertices respectively. A complete bipartite graph is built upon the two sets of vertices, and the weight of each edge is the similarity between corresponding vertices calculated by the person search network. Then the Kuhn-Munkres (K-M) algorithm (Kuhn 1955; Munkres 1957) is exploited to discovery the optimal matching with maximum weight. In this matching, the person connected with the querier is taken as top-1 result.

The contributions of this paper are three-fold:

- We notice that the performance of previous end-to-end framework is limited by inferior features and formulate a Sequential End-to-end Network (SeqNet) to refine them.
- To make full use of context information, we propose a Context Bipartite Graph Matching algorithm to perform more robust matching.
- Our method outperforms all other state-of-the-art ones on the two widely used benchmarks CUHK-SYSU (Xiao et al. 2017) and PRW (Zheng et al. 2017). Moreover, our method can be integrated into the existing end-to-end framework easily.

Related Work

Person Search

Person search has raised a lot of interest in computer vision community since the publication of two large scale datasets, CUHK-SYSU (Xiao et al. 2017) and PRW (Zheng et al. 2017). It's a straightforward solution to tackle the problem with a pedestrian detector and a re-ID descriptor sequentially. Zheng et al. make a systematic evaluation on various detectors and descriptors, and propose a re-weighting algorithm adjusting the matching similarity to suppress the false positive detections (Zheng et al. 2017). Lan, Zhu, and Gong point out the performance of person search is limited by the multi-scale matching, and formulates a Cross-Level Semantic Alignment (CLSA) method capable of learning more discriminative identity representations (Lan, Zhu, and Gong 2018). Chen et al. first reveal the inherent optimization conflict between the pedestrian detection and person re-ID, and present a Mask-Guided Two-Stream (MGTS) method to eliminate the conflict (Chen et al. 2018). Han et al. introduce a RoI transform layer to jointly optimize the detection and re-ID models (Han et al. 2019). Wang et al. notice the consistency requirements between the two subtasks in person search, and adopt a Task-Consistent Two-Stage (TCTS) framework to solve the inconsistency existing in previous works (Wang et al. 2020). Dong et al. propose a Instance Guided Proposal Network (IGPN) to reduce the number of proposals to relieve the burden of re-ID (Dong et al. 2020b).

Besides the two-stage framework, the faster and simpler end-to-end methods based on Faster R-CNN are also popular. Xiao et al. design the first end-to-end person search network, which is trained with standard Faster R-CNN losses and their proposed Online Instance Matching (OIM) loss (Xiao et al. 2017). Xiao et al. introduce center loss to increase the intra-class compactness of feature representations (Xiao et al. 2019). Instead of generating BBoxes for all people in the image, Liu et al. propose to recursively shrinking the search area under the guidance of the query (Liu et al. 2017). Chang et al. adopt a similar idea and first introduce the deep reinforcement learning into person search framework (Chang et al. 2018). Yan et al. build a graph model to exploit context information as an complementary cue for person matching (Yan et al. 2019). Munjal et al. propose a query-guided region proposal network (QRPN) to produce query-relevant proposals, and a query-guided similarity subnetwork (QSimNet) to learn a query-guided re-ID score (Munjal et al. 2019). Chen et al. propose a Hierarchical Online Instance Matching (HOIM) loss which exploits the hierarchical relationship between detection and re-ID to guide the feature learning of their network (Di Chen14 et al. 2020). Dong et al. design a Bi-directional Interaction Network (BINet) to remove rebundent context information outside BBoxes (Dong et al. 2020a). To reconcile the contradictory goals of the two subtasks, Chen et al. present a novel

approach called Norm-Aware Embedding (NAE) to disentangle the person embedding into norm and angle for detection and re-ID respectively (Chen et al. 2020).

Multi-stage Faster R-CNN

Some researchers extend Faster R-CNN to a multi-stage fashion. Gidaris and Komodakis propose a post-processing step that the network iterate several times in inference stage to achieve better localization performance (Gidaris and Komodakis 2015, 2016). Cai and Vasconcelos design a cascade framework containing a sequence of detectors trained with increasing IoU thresholds to be sequentially more selective against close false positives (Cai and Vasconcelos 2018). Inspired by them, our model is designed as a multi-stage framework to introduce the sequential process into end-to-end person search network.

Method

In this section, we first revisit the end-to-end person search network, then discuss its shortcoming. Next, we describe our proposed Sequential End-to-end Network (SeqNet). Finally, we formulate a Context Bipartite Graph Matching (CBGM) algorithm to utilize the context information.

End-to-end Network for Person Search

We take the multi-task network NAE (Chen et al. 2020) as our baseline. The overview of this baseline is illustrated in Figure 3 (a). It adopt ResNet50 (He et al. 2016) as the backbone network. Specifically, res1∼res4 are taken as the stem network to extract the 1024-channel stem feature maps of the image. A Region Proposal Network (RPN) is built upon these feature maps to generate region proposals. After NMS, we keep 128 proposals, and exploit RoI-Align to pool a $1024 \times 14 \times 14$ region for each of them. Next these regions are fed into res5 to extract 2048-dim features, which are then mapped to 256-dim. It uses these 2048-dim features to calculate regressors and 256-dim features to perform classification and re-ID tasks. The Norm-Aware-Embedding is designed to supervise the classification and re-ID branches and the Smooth-L_1-Loss (Girshick 2015) is adopted to supervise the regression branch.

Problems of the End-to-end Framework

As aforementioned, the baseline suffers from the inferior features. To investigate its influence, we train the baseline model and report the results under two evaluation settings in Table 1. The original setting is denoted by *parallelization*. In the second setting called *serialization*, the network will iterate twice to solve the detection and re-ID in turn. The first iteration will output detected BBoxes. Then we exploit RoI-Align to pool a fixed size region for each BBox and feed them into res5 to extract re-ID features. In this way, superior BBoxes can be obtained before re-ID stage. Table 1 shows the mAP of re-ID is increased by 0.75% on CUHK-SYSU, 1.12% on PRW. This demonstrates the re-ID ability of the network is greatly limited by the inferior features of proposals. We also notice that the detection performance has a slight decline. This is caused by the inconsistency between the training and test phase, *i.e.*, the network is trained by the proposals generated by RPN, but tested by the detected BBoxes. Therefore it is necessary to introduce serialization into model training, rather than just in test phase.

Method	Detection		re-ID	
	Recall	AP	mAP	top-1
CUHK-SYSU				
parallelization	92.6	86.6	91.7	92.8
serialization	90.9	85.7	**92.5**	**93.7**
PRW				
parallelization	93.8	88.7	43.6	80.0
serialization	93.7	89.4	**44.7**	**80.8**

Table 1: Influence of inferior features on the performance on CUHK-SYSU and PRW datasets. We separate the person search into detection and re-ID, and evaluate their performance individually. The bold font represents the best result. Most experiment results will be presented in this form.

Proposed Sequential End-to-end Network

The overview of our model is shown in Figure 3 (b). It consists of two head networks to solve person detection and person re-ID respectively. The first standard Faster R-CNN head is employed to generate accurate BBoxes. The second unmodified baseline head is applied to further fine-tune these BBoxes and extract their discriminative features.

The main idea is to exploit Faster R-CNN as a stronger RPN to provide fewer but more accurate candidate BBoxes. These high-quality BBoxes lead to more discriminative embeddings.

Training During training phase, these two heads are trained with 0.5 IoU threshold to distinguish positive and negtive samples, and the feature learning is supervised by the following 5 losses.

- L_{reg_1}/L_{reg_2}: The regression loss of the first/second head. N_p is the number of positive samples, r_i is the calculated regressor of i-th positive sample, $\triangle_i$ is the corresponding ground truth regressor, and L_{loc} is the Smooth-L_1-Loss.

$$L_{reg} = \frac{1}{N_p} \sum_{i=1}^{N_p} L_{loc}(r_i, \triangle_i) \tag{1}$$

- L_{cls_1}: The classification loss of the first head. N is the number of samples, p_i is the predicted classification probability of i-th sample, and c_i is the ground truth label.

$$L_{cls_1} = -\frac{1}{N} \sum_{i=1}^{N} c_i log(p_i) \tag{2}$$

- L_{cls_2}, L_{reid}: The classification and re-ID losses of the second head. It is calculated by the Norm-Aware-Embedding $L_{nae}(.)$. f is the extracted 256-dim features.

$$L_{cls_2}, L_{reid} = L_{nae}(f) \tag{3}$$

The overall learning objective function is given as:

$$L = \lambda_1 L_{reg1} + \lambda_2 L_{cls1} + \lambda_3 L_{reg2} + \lambda_4 L_{cls2} + \lambda_5 L_{reid} \tag{4}$$

λ_1 is set to 10, and the others are 1.

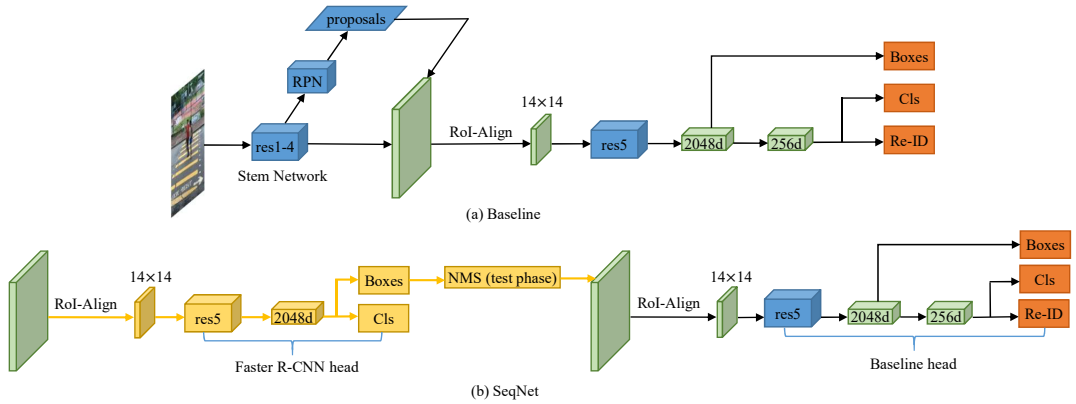

Figure 3: (a). Baseline (b). Our Sequential End-to-end Network, in which yellow parts are modifications and NMS only be applied in inference stage. The structure before RoI-Align is the same as baseline, so it is not shown here for simplification.

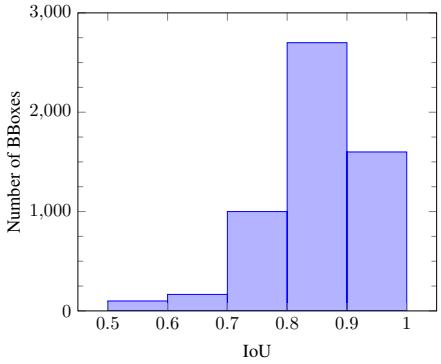

Figure 4: The IoU statistics of the BBoxes of labeled pedestrians detected by the first head in test phase.

Inference In inference stage, NMS is applied to remove redundant BBoxes before re-ID stage. In this way, the inference speed will be greatly accelerated.

First Classification Score (FCS) Figure 4 shows that there are a lot of detected BBoxes with IoU > 0.8 in test phase. The second head is trained with 0.5 IoU threshold, so it may fail to classify these high IoU samples correctly. Hence we take the more reliable classification scores predicted by the first head as output.

Discussion Our SeqNet shares the similar structure with previous works (Gidaris and Komodakis 2015, 2016; Cai and Vasconcelos 2018), but our method differs from them significantly from the following aspects:

- **Motivation** Previous multi-stage Faster R-CNN is proposed to achieve better detection performance. However, our method aims at solving the detection and re-ID sequentially with a jointly optimized network to extract more discriminative features.
- **Efficiency** Our SeqNet owns an extra NMS, which ensures the efficiency in test phase. In contrast, each head of their networks needs to handle all BBoxes.

Context Bipartite Graph Matching

In this section, we present a novel Context Bipartite Graph Matching (CBGM) algorithm used in test phase to integrate context information into the matching process.

Traditional person search task can be seen as a single-point matching strategy, which takes the person most similar to the querier in the gallery image as the search result. But it may fail when there are multiple people with very similar appearances in the gallery image. We extend it to a multi-point matching strategy, which matches both the querier and its surrounding people with all the detected pedestrians in the gallery image. In this way, when the single-point matching strategy fails, as long as the surrounding people can be correctly matched, the query person can still be identified.

Taking Figure 2 for example, we define the following symbols.

- Q/G: The query/gallery image (the upper/lower image).
- q: The query person in Q (the blue box, *i.e.*, person (a)).
- V: All people in image ($V_G = \{(c), (d)\}$).
- $sim(p_1, p_2)$: The cosine similarity between person p_1 and p_2 calculated by extracted features.
- $SIM(q, G)$: The similarity between q and G. It is defined as the maximum value among these similarities between q and all people in G.

$$SIM(q, G) = \max_{p \in V_G} sim(q, p) \tag{5}$$

In graph theory, a *matching* $M = (\mathbb{V}, \mathbb{E})$ in an undirected graph is a set of edges without common vertices. $\mathbb{V}$ is the set of vertices and $\mathbb{E}$ is the set of edges. We further define the following concepts.

- $weight(e_i, e_j)$: The weight of the edge $(e_i, e_j) \in \mathbb{E}$.
- $weight(M)$: The weight of matching M. It is defined as the sum of the weights of all edges.

$$weight(M) = \sum_{(e_i, e_j) \in \mathbb{E}} weight(e_i, e_j) \tag{6}$$

- $C(M)$: The confidence of matching M. It is defined as the maximum value among all weights.

$$C(M) = \max_{(e_i,e_j)\in\mathbb{E}} weight(e_i, e_j) \tag{7}$$

Based on these two sets of vertices V_Q and V_G, we firstly build a complete bipartite graph $G = (\mathbb{V}, \mathbb{E})$, in which $\mathbb{V} = V_Q \cup V_G$. The graph has the following properties:

- For every two vertices $v_1 \in V_Q$ and $v_2 \in V_G$, (v_1, v_2) is an edge in $\mathbb{E}$.
- No edge has both endpoints in the same set of vertices.
- For each edge $(e_i, e_j) \in \mathbb{E}$, its weight is the similarity of corresponding vertices, *i.e.*, $weight(e_i, e_j) = sim(e_i, e_j)$.

Then the Kuhn-Munkres (K-M) algorithm (Kuhn 1955; Munkres 1957) is exploited to find the optimal matching with largest $weight(M)$. In Figure 2, the matching is $(a) \leftrightarrow (c), (b) \leftrightarrow (d)$, and the query person (a) can be correctly matched with the ground truth (c).

The proposed Context Bipartite Graph Matching (CBGM) algorithm is described in Algorithm 1. We rank all the gallery images in descending order by $SIM(q, G)$, and remain the top-k_1 to be processed. In this way, most gallery images in which q does not appear can be removed. Additionally, excessive context information may bring noise. Therefore we only regard the people with top-k_2 detection confidence in the query image as context information. After the optimal matching M is found, $C(M)$ is taken as the similarity between q and its matched person.

Algorithm 1 CBGM

Input:
Query image, Q
Query person, $q \in V_Q$
Gallery images, $S = \{G_1, G_2, ...\}$
Number of processed gallery images, k_1
Maximum context, k_2

Output:
Most similar person in each gallery image
Similarities between q and these most similar people

Rank S in descending order by $SIM(q, G)$
Remain top-k_1 gallery images, $S = \{G_1, G_2, ..., G_{k_1}\}$
Rank V_Q in descending order by detection confidence
Remain top-k_2 people, $V_Q = \{q_1, q_2, ..., q_{k_2}\}$
Set $people, sims$ to empty list
for each $G \in S$ **do**
  Based on V_Q and V_G, build a complete bipartite graph $G = (\mathbb{V}, \mathbb{E})$
  Exploit K-M algorithm to find the optimal matching M with largest weight
  for each edge (e_i, e_j) of M **do**
    if $e_i = q$ **then**
      Insert e_j into $people$
      Insert $C(M)$ into $sims$
      break
    end if
  end for
end for
return $people, sims$

Experiments

In this section, we first introduce the datasets and evaluation protocols. Then we describe the implementation details, followed by ablation studies on the efficacy of each component. Finally, we compare our method with state-of-the-art ones.

Datasets and Evaluation Protocol

CUHK-SYSU CUHK-SYSU (Xiao et al. 2017) is a large scale person search dataset containing 18,184 scene images and 96,143 annotated BBoxes, which are collected from two sources: street snap and movie. All people are divided into 8,432 labeled identities and other unknown ones. The training set contains 11,206 images and 5,532 different identities. The test set contains 6,978 images and 2,900 query people. The training and test sets have no overlap on images and query people. For each query, different gallery sizes from 50 to 4000 are pre-defined to evaluate the search performance. If not specify, gallery size of 100 is used by default.

PRW PRW is another widely used dataset (Zheng et al. 2017) containing 11,816 video frames captured by 6 cameras in Tsinghua university. 34,304 BBoxes are annotated manually. Similar to CUHK-SYSU, all people are divided into labeled and unlabeled identities. The training set contains 5,704 images and 482 different people, while the test set includes 6,112 images and 2,057 query people. For each query, the gallery is the whole test set, *i.e.*, the gallery size is 6112.

Evaluation Protocol Following the settings in previous works (Munjal et al. 2019; Chen et al. 2020), the Cumulative Matching Characteristic (CMC) and the mean Averaged Precision (mAP) are adopted as the performance metrics. The formal is widely used in person re-ID, and the latter is inspired by object detection task. The higher the two metrics, the better the performance.

Implementation Details

We implement our model with PyTorch (Paszke et al. 2017) and run all experiments on one NVIDIA Tesla V100 GPU. We adopt ResNet50 (He et al. 2016) pretrained on the ImageNet (Deng et al. 2009) as the backbone network. During training, batch size is 5 and each image is resized to 900×1500 pixels. Our model is optimized by Stochastic Gradient Descent (SGD) for 20 epochs (18 epochs for PRW) with initial learning rate of 0.003 which is warmed up during the first epoch and decreased by 10 at the 16-th epoch. The momentum and weight decay of SGD are set to 0.9 and 5×10^{-4} individually. For CUHK-SYSU/PRW, the circular queue size of OIM is set to 5000/500. At test time, NMS with 0.4/0.5 threshold is used to remove redundant boxes detected by the first/second head.

Ablation Study

In this section, we perform several analytical experiments on CUHK-SYSU to better understand our proposed method.

Detector	Recall	AP	Re-identifier	mAP	top-1
NAE	92.6	86.6	NAE	92.5	93.7
			SeqNet	93.1	94.0
SeqNet	92.1	89.2	NAE	93.3	93.8
			SeqNet	93.8	94.5
GT	100	100	NAE	94.1	94.6
			SeqNet	94.6	95.3

Table 2: Analytical experiment results with different detectors and re-identifiers on CUHK-SYSU.

Different detectors and re-identifiers We first explore whether the improvement brought by SeqNet comes from better detection or more discriminative features. We separate the person search task into two stages: detection stage with different detectors and re-ID stage with different re-identifiers. When using NAE re-identifier, we remove its RPN module and set the proposals manually to the BBoxes detected by the specified detector (*e.g.* SeqNet). In particular, NAE detector + NAE re-identifier is equivalent to the *serialization* mentioned in last section. The results are summarized in Table 2, from which we can draw the following conclusions:

- **The detection of SeqNet is better** We can see from the second column that SeqNet (Recall: 92.1, AP: 89.2) achieves better detection than NAE (Recall: 92.6, AP: 86.6) in overall. It is mainly because that each head (RPN head/Faster R-CNN head/baseline head) of SeqNet will perform regression to BBoxes, which makes our model more selective against false positives.
- **SeqNet is more discriminative for re-ID** When using NAE detector, the mAP and top-1 accuracy of SeqNet outperform that of NAE by 0.6% and 0.3% respectively. Similar improvement (mAP $\uparrow$ 0.5%, top-1 $\uparrow$ 0.7%) can be observed when using SeqNet detector. This demonstrates our SeqNet can extract more discriminative features with the same detection ability. This is caused by the inconsistency of NAE, *i.e.*, trained by low-quality proposals but testd by high-quality detected BBoxes. In contrast, the baseline head of SeqNet is trained with detected BBoxes, which makes it more suitable for test scenario.
- **Detection is not the performance bottleneck** If ground truth BBoxes are adopted as detection results, the mAP of NAE can be increased by 2.4%, while SeqNet can only be increased by 0.8%. It indicates that SeqNet gains very little from better detection, and future research should focus on how to achieve a better re-ID.

FCS and NMS SeqNet has two key components: FCS to improve the classification ability, NMS to accelerate the inference speed. The upper block of Table 3 shows that FCS greatly improves the detection (Recall: 91.5→92.7, AP: 86.7→89.7), which leads to a better re-ID (mAP: 93.1→93.8, top-1: 94.0→94.5). In addition, although NMS slightly reduces detection performance (Recall: 92.7→92.1, AP: 89.7→89.2), it does not affect re-ID and increases the FPS (processed Frames Per Second) from 7.4 to 11.5.

The multi-task framework of baseline head The lower block of Table 3 reports the impact of each task of the baseline head. Since the detection of Faster R-CNN head is strong enough, the regression and classification branches of baseline head will not have much impact on the overall detection, but they will facilitate the re-ID branch to learn more discriminative features. We can observe that neither regression nor classification branch alone can achieve the best performance, suggesting that the two are complementary.

FCS	NMS	Re-ID	Cls	Reg	Detection		Re-ID		FPS
					Recall	AP	mAP	top-1	
		✓	✓	✓	91.5	86.7	93.1	94.0	7.4
✓		✓	✓	✓	92.7	89.7	93.8	94.5	7.4
	✓	✓	✓	✓	89.1	86.3	93.4	94.4	11.5
✓	✓	✓	✓	✓	92.1	89.2	93.8	94.5	11.5
✓	✓	✓			92.6	89.5	93.0	93.8	-
✓	✓	✓		✓	92.8	89.5	93.3	94.1	-
✓	✓	✓	✓		92.4	89.4	93.4	94.2	-

Table 3: Influence of different components on accuracy and speed. The ablation study about FCS and NMS is in the upper block. The multi-task framework of baseline head is discussed in the lower block.

k_2 \ k_1	10	20	30	40	50
CUHK-SYSU					
2	94.9	94.8	94.8	94.8	94.8
3	**95.2**	95.1	95.0	95.0	95.0
4	95.1	94.9	94.7	94.6	94.6
5	95.0	94.7	94.6	94.5	94.4
6	95.0	94.8	94.7	94.5	94.5
PRW					
2	66.0	66.6	66.7	66.5	66.5
3	66.4	67.0	67.3	67.2	67.1
4	66.6	67.2	**67.6**	67.4	67.1
5	66.6	67.2	67.5	67.3	67.0
6	66.6	67.2	67.5	67.5	67.2

Table 4: The performance on CUHK-SYSU (the upper block) and PRW (the lower block) datasets with different k_1 and k_2 of CBGM. We evaluate the performance by $\frac{mAP+top-1}{2}$.

Different k_1 and k_2 of CBGM We evaluate the performance with different k_1 and k_2 of CBGM. Figure 4 shows that CBGM is robust to these two parameters. On CUHK-SYSU, $k_1/k_2 = 10/3$ achieves the best performance, while on PRW, $k_1/k_2 = 30/4$ is the best choice. This is because that the gallery size of PRW (6112) is much larger than that of CUHK-SYSU (100). Therefore, k_1 and k_2 needs to be larger to capture more context information for precise search.

Efficiency of CBGM Table 5 reports the average time to search a query person under different gallery sizes. For CUHK-SYSU dataset, after sorting all gallery images in descending order by $SIM(q, G)$, CBGM is applied to the top-10 gallery images. Table 5 shows that the additional computation brought by CBGM is fixed (about 2ms) and light. The larger the gallery size, the smaller the impact of CBGM on speed.

Integrated into another method To verify the universality of our method, we integrate SeqNet into the existing end-to-end framework. We choose the widely studied OIM

Method	Gallery size				
	100	**500**	**1000**	**2000**	**4000**
SeqNet	347	366	390	439	541
SeqNet+CBGM	349	368	392	441	542

Table 5: The average time to search a query person under different gallery sizes on CUHK-SYSU. The unit is milliseconds.

	Method	CUHK-SYSU		PRW	
		mAP	**top-1**	**mAP**	**top-1**
two-stage	DPM(Girshick et al. 2015)	-	-	20.5	48.3
	MGTS(Chen et al. 2018)	83.0	83.7	32.6	72.1
	CLSA(Lan, Zhu, and Gong 2018)	87.2	88.5	38.7	65.0
	RDLR(Han et al. 2019)	93.0	94.2	42.9	70.2
	IGPN(Dong et al. 2020b)	90.3	91.4	47.2	87.0
	TCTS(Wang et al. 2020)	93.9	95.1	46.8	87.5
end-to-end	OIM(Xiao et al. 2017)	75.5	78.7	21.3	49.9
	IAN(Xiao et al. 2019)	76.3	80.1	23.0	61.9
	NPSM(Liu et al. 2017)	77.9	81.2	24.2	53.1
	RCAA(Chang et al. 2018)	79.3	81.3	-	-
	CTXGraph(Yan et al. 2019)	84.1	86.5	33.4	73.6
	QEEPS(Munjal et al. 2019)	88.9	89.1	37.1	76.7
	HOIM(Di Chen14 et al. 2020)	89.7	90.8	39.8	80.4
	BINet(Dong et al. 2020a)	90.0	90.7	45.3	81.7
	NAE(Chen et al. 2020)	91.5	92.4	43.3	80.9
	NAE+(Chen et al. 2020)	92.1	92.9	44.0	81.1
	OIM(ours)	87.1	88.5	34.0	75.9
	OIM+SeqNet(ours)	93.4	94.1	45.8	81.7
	OIM+SeqNet+CBGM(ours)	94.3	95.0	46.6	84.9
	NAE+SeqNet(ours)	93.8	94.6	46.7	83.4
	NAE+SeqNet+CBGM(ours)	**94.8**	**95.7**	**47.6**	**87.6**

Table 6: Comparison of mAP and top-1 accuracy with the state-of-the-art methods on CUHK-SYSU and PRW. Our models are shown in italics.

(Xiao et al. 2017) as the base network and its implementation is the same as in (Chen et al. 2020). As shown in Table 6, SeqNet improves the mAP of OIM by 6.3% and 11.8% on CUHK-SYSU and PRW benchmarks respectively, which demonstrates that our method is insensitive to base network. Particularly, OIM+SeqNet+CBGM further achieves 94.3 of mAP and 95.0 of top-1 accuracy, outperforming all other competitors. It indicates the great potential of our method.

Comparison with the State-of-the-art Methods

In this section, we compare our method with the state-of-the-art models on CUHK-SYSU and PRW.

CUHK-SYSU Table 6 shows that both the mAP and top-1 accuracy of our method are higher than other competitors. Compared with the state-of-the-art two-stage model TCTS, our NAE+SeqNet+CBGM outperforms it by 0.9% and 0.6% w.r.t mAP and top-1 accuracy though it adopts more tricks, *e.g.*, label smooth, random erasing (Zhong et al. 2020), and triplet loss (Hermans, Beyer, and Leibe 2017). This demonstrates the effectiveness of solving detection and re-ID jointly, which can avoid sub-optimal solution.

We also compare these methods under different gallery sizes. Figure 5 shows that the mAP of all algorithms decreases monotonically with the increase of gallery size, which indicates the difficulty of locating a target person in a large search scope. We can observe that our method still ranks best at all gallery sizes.

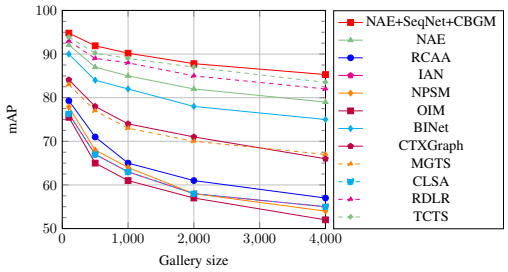

Figure 5: The mAP under different gallery sizes. The dashed lines represent two-stage methods and the solid lines represent end-to-end ones.

GPU(TFLOPs)	MGTS	QEEPS	NAE	NAE+	SeqNet
K80(4.1)	1269	-	663	606	-
P6000(12.6)	-	300	-	-	-
P40(11.8)	-	-	158	161	-
V100(14.1)	-	-	83	98	86

Table 7: Comparison of running time on different GPUs. The unit is milliseconds.

PRW The right column of Table 6 summarizes the results on PRW dataset. Our NAE+SeqNet+CBGM still surpasses the others. PRW has fewer training data than CUHK-SYSU, which indicates that our method is robust and effective for a relatively small dataset.

Runtime Comparison We compare the speed of different models, and show the Tera-Floating Point Operation per-second (TFLOPs) for each GPU for fair comparison. Our SeqNet is implemented in PyTorch, and images are resized to 900×1500 pixels, which is the same as MGTS and QEEPS. Table 7 shows that our method is around 2 times faster than QEEPS and MGTS. Finally, our SeqNet costs 86 milliseconds per-frame on a V100 GPU, which is only a bit slower than NAE. The fast speed reveals the great potential of SeqNet to real-world applications.

Conclusion

In this paper, we notice the performance of previous end-to-end framework is limited by inferior features. To address the issue, we propose a Sequential End-to-end Network to solve the detection and re-ID in turn. Besides, we design a Context Bipartite Graph Matching algorithm to exploit context information as a complement to individual feature. Extensive experiments demonstrate that our method can significantly improve the performance of previous end-to-end models at an acceptable time cost.

Acknowledgements

This research was supported in part by the National Nature Science Foundation of China (Grant Nos. 61976158 and 61673301).

References

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6154–6162.

Chang, X.; Huang, P.-Y.; Shen, Y.-D.; Liang, X.; Yang, Y.; and Hauptmann, A. G. 2018. RCAA: Relational context-aware agents for person search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 84–100.

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

Chen, D.; Zhang, S.; Yang, J.; and Schiele, B. 2020. Norm-Aware Embedding for Efficient Person Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12615–12624.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Di Chen14, S. Z.; Ouyang, W.; Yang, J.; and Schiele, B. 2020. Hierarchical Online Instance Matching for Person Search .

Dong, W.; Zhang, Z.; Song, C.; and Tan, T. 2020a. Bi-Directional Interaction Network for Person Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2839–2848.

Dong, W.; Zhang, Z.; Song, C.; and Tan, T. 2020b. Instance Guided Proposal Network for Person Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2585–2594.

Fu, Y.; Wei, Y.; Zhou, Y.; Shi, H.; Huang, G.; Wang, X.; Yao, Z.; and Huang, T. 2019. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8295–8302.

Gidaris, S.; and Komodakis, N. 2015. Object detection via a multi-region and semantic segmentation-aware cnn model. In *Proceedings of the IEEE international conference on computer vision*, 1134–1142.

Gidaris, S.; and Komodakis, N. 2016. Attend refine repeat: Active box proposal generation via in-out localization. *arXiv preprint arXiv:1606.04446* .

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587.

Girshick, R.; Iandola, F.; Darrell, T.; and Malik, J. 2015. Deformable part models are convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 437–446.

Han, C.; Ye, J.; Zhong, Y.; Tan, X.; Zhang, C.; Gao, C.; and Sang, N. 2019. Re-id driven localization refinement for person search. In *Proceedings of the IEEE International Conference on Computer Vision*, 9814–9823.

Hao, Y.; Wang, N.; Li, J.; and Gao, X. 2019. HSME: Hypersphere manifold embedding for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8385–8392.

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

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* .

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly* 2(1-2): 83–97.

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

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

Munjal, B.; Amin, S.; Tombari, F.; and Galasso, F. 2019. Query-guided end-to-end person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 811–820.

Munkres, J. 1957. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics* 5(1): 32–38.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch .

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Wang, C.; Ma, B.; Chang, H.; Shan, S.; and Chen, X. 2020. TCTS: A Task-Consistent Two-Stage Framework for Person Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11952–11961.

Wang, G.; Lai, J.; Huang, P.; and Xie, X. 2019. Spatial-temporal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8933–8940.

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

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

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

Yan, Y.; Zhang, Q.; Ni, B.; Zhang, W.; Xu, M.; and Yang, X. 2019. Learning context graph for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2158–2167.

Yang, Y.; Wen, L.; Lyu, S.; and Li, S. Z. 2017. Unsupervised learning of multi-level descriptors for person re-identification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4306–4312.

Zhao, C.; Lv, X.; Zhang, Z.; Zuo, W.; Wu, J.; and Miao, D. 2020. Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification. *IEEE Transactions on Multimedia* .

Zhao, C.; Wang, X.; Chen, Y.; Gao, C.; Zuo, W.; and Miao, D. 2017. Consistent iterative multi-view transfer learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 1087–1094.

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

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random Erasing Data Augmentation. In *AAAI*, 13001–13008.